A Section 508–conformant HTML version of this article is available at https://doi.org/10.1289/EHP3264.

# Best Practices for QSAR Model Reporting: Physical and Chemical Properties, Ecotoxicity, Environmental Fate, Human Health, and Toxicokinetics Endpoints

*Geven Piir,[1] Iiris Kahn,[2] Alfonso T. García-Sosa,[1] Sulev Sild,[1] Priit Ahte,[2] and Uko Maran[1]*

[1]Institute of Chemistry, University of Tartu, Tartu, Estonia
[2]Department of Chemistry and Biotechnology, Tallinn University of Technology, Tallinn, Estonia

**Background:** Quantitative and qualitative structure–activity relationships (QSARs) have been used to understand chemical behavior for almost a century. The main source of QSAR models is the scientific literature, but the open question is how well these models are documented.

**Objectives:** The main aim of this study was to critically analyze the publication practices of QSARs with regard to transparency, potential reproducibility, and independent verification. The focus was on the level of technical completeness of the published QSARs.

**Methods:** A total of 1,533 QSAR articles reporting 79 individual endpoints, mostly in environmental and health science, were reviewed. The QSAR parameters required for technical completeness were grouped into five categories: chemical structures, experimental endpoint values, descriptor values, mathematical representation of the model, and predicted endpoint values. The data were summarized and discussed using Circos plots.

**Results:** Altogether, 42.5% of the reviewed articles were found to be potentially reproducible. The potential reproducibility for different endpoint groups varied; the respective rates were 39% for physical and chemical properties, 52% for ecotoxicity, 56% for environmental fate, 30% for human health, and 32% for toxicokinetics. The reproducibility of QSARs is discussed and placed in the context of the reproducibility of the experimental methods. Included are 65 references to open QSAR datasets as examples of models restored from scientific articles.

**Discussion:** Strikingly poor documentation of QSARs was observed, which reduces the transparency, availability, and consequently, the application of research results in scientific, industrial, and regulatory areas. A list of the components needed to ensure the best practices for QSAR reporting is provided, allowing long-term use and preservation of the models. This list also allows an assessment of the reproducibility of models by interested parties such as journal editors, reviewers, regulators, evaluators, and potential users. https://doi.org/10.1289/EHP3264

## Introduction

Quantitative and qualitative structure–activity relationships, QSARs, is a modeling approach that has been an essential way of thinking and toolbox for more than a century. QSARs have been used in many areas of natural science to gather information and create new knowledge by linking molecular or material structures to chemistry-driven phenomena. QSAR has its mechanistic roots in physical organic chemistry and has provided a wealth of knowledge on chemical reactivity (Hansch et al. 1991). Equally prominent are landmarks in the fields of medicinal chemistry (Hansch et al. 1996, 2002; Cherkasov et al. 2014), drug design (Seddon et al. 2012), and predictive and computational toxicology (Dearden 2016, 2017); these landmarks have facilitated the design of novel bioactive compounds (see Berhanu et al. 2012; Boyd and March 2006 for an extensive list of examples) and have been used to estimate the environmental safety of existing and new chemical entities (Price and Watkins 2003; Katritzky et al. 2010). The vitality of QSAR is also evident from its success in predictive modeling of technologically relevant properties (Katritzky et al. 2000) and in exploratory applications, such as materials (Le et al. 2012; Käärik et al. 2018), ionic liquids (Das and Roy 2013), and chemical mixtures (Muratov et al. 2012). QSAR has been found to be invaluable in various decision-support scenarios in the pharmaceutical industry (Cumming et al. 2013), in regulatory use (Cronin et al. 2003b; Cronin et al. 2003a; Benfenati et al. 2007; Kruhlak et al. 2007; Gallegos Salinger et al. 2007; Tsakovska et al. 2007; OECD 2007), and, recently, in the systematic analysis of adverse outcome pathways of chemicals (Patlewicz and Fitzpatrick 2016). QSAR continuously faces new challenges. For example, in the past decade, researchers have implemented QSAR methodological solutions to describe and predict the properties of nanostructures and nanomaterials, as well as to explain the processes behind these properties (Winkler et al. 2014). However, progress in this area has been limited by the quality of the data available for modeling, and the field has largely remained in the phase of searching for methodological solutions, mainly how to quantify structure for modeling (Burello and Worth 2011; Tantra et al. 2015). The many roles of QSAR as a scientific methodology have made it a unique approach for gaining new knowledge (Fujita and Winkler 2016).

Classical QSARs were traditionally developed in the form of multilinear regression (MLR). The evolution of machine learning methods and their application to explain chemical phenomena allowed QSARs to expand beyond their original frames. In fact, the mathematical representation of QSAR models today is often more diverse and complex. Algorithms such as k-nearest neighbors (k-NN), linear discriminant analysis (LDA), decision trees (DT), random forests (RF), artificial neural networks (ANN), support vector machines (SVM), naïve Bayes models, ensemble models, and others are being used more frequently. These developments and the expanding experience in building QSAR models have prompted various discussions in the literature about the best practices for QSAR model development (Gedeck et al. 2010; Scior et al. 2009; Tropsha 2010; Martin et al. 2012). The growing use of QSARs in decision-support systems has led to studies and discussions on the validation of models, their applicability, and the uncertainty of their predictions (Eriksson et al. 2003; Tropsha et al. 2003; Netzeva et al. 2005; Tetko et al. 2006; Gramatica 2007; Chirico and Gramatica 2011; Alexander et al. 2015; Golbraikh and Tropsha 2002). In recent years, the increasing amount of chemical data from the wave of big data in chemistry (Tetko et al. 2016) introduced the importance of data curation in

Address correspondence to U. Maran, Institute of Chemistry, University of Tartu, 14A Ravila St., Tartu 50411, Estonia. Telephone: +3727375254. Fax: +3727375264. Email: uko.maran@ut.ee

Supplemental Material is available online (https://doi.org/10.1289/EHP3264).

The authors declare they have no actual or potential competing financial interests.

Received 19 December 2017; Revised 19 October 2018; Accepted 7 November 2018; Published 18 December 2018.

QSAR modeling (Fourches et al. 2010; Clark and Waldman 2012; Ruusmann and Maran 2013; Fourches et al. 2016). QSAR, like any good scientific method, faces challenges (Combes 2001). The scientific literature has reported debates about and analyses of the selection of proper model(s) (Johnson 2007), the validation of models (Huang and Fan 2011), and the predictivity or reliability of models (Benigni and Bossa 2008a, 2008b), and has provided examples of developer errors (Dearden et al. 2009). Sometimes, insufficient information about a model can lead to unrealistically high expectations or to improper use of the model (Stouch et al. 2003). Despite these challenges, if used in a proper manner, QSAR functions as an important and very often primary tool in studies of chemical phenomena, especially when fine-tuning and understanding of chemical behavior are needed (Doweyk 2008).

QSARs are typically made available in scientific articles; some models have found their way into commercial software (Nicolotti et al. 2014) or, in fewer cases, into public solutions and, more recently, onto the web without paywalls (Tetko et al. 2017). Following the last-century milestone achieved in 1964 by Hansch and Fujita (1964) and the progress that has been made in computational tools and software applications during the past few decades, QSAR methods have been extensively used. According to the Web of Science Core Collection (WOS), from the year 2007 onwards, over 1,000 articles per year that refer to the terms "QSAR" or "QSPR (quantitative structure—property relationship)" (Figure 1) have been published in the peer-reviewed literature. The Scopus literature database (SCOPUS) reached over 1,000 "QSAR" or "QSPR" articles in the year 2008 (Figure 1). The actual number of articles with QSAR models is rather difficult to estimate from this search because some articles refer to already existing and published models, and not all articles that contain QSARs appear in a straightforward way in search results. Nevertheless, the number of QSAR articles is steadily increasing due to the wider availability of experimental data on chemicals and the growth of statistical and machine learning methods. The scientific literature is unarguably the main source of QSAR models, and this brings us to the subject matter of this review—how well published models are documented to allow their reproduction and extended practical use.

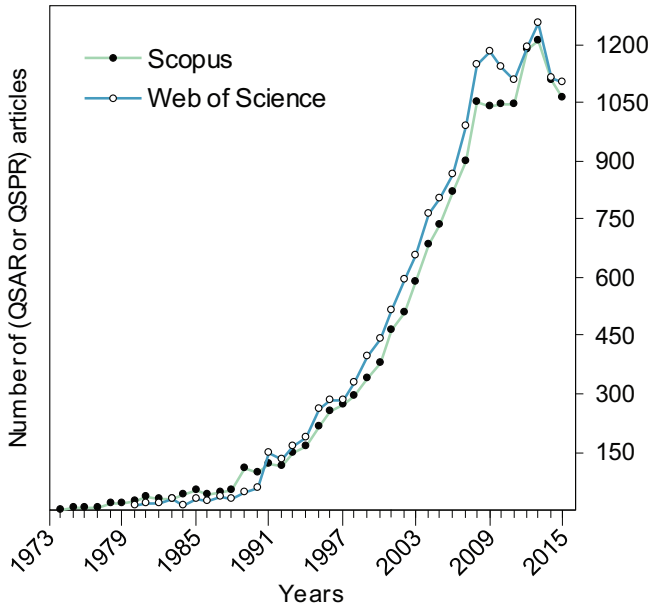


**Figure 1.** Number of quantitative and qualitative structure–activity relationships (QSAR) and QSPR articles in the peer-reviewed literature during the years 1974–2015 according to the Web of Science Core Collection and Scopus literature databases (search performed on April 3, 2016).

The reproducibility of QSAR models is a main concern for their acceptance in regulatory use; this has been stressed by several authors (Dearden et al. 2009; Hartung et al. 2004). Moreover, the reproducibility of research is a fundamental assumption in science; it allows independent verification of scientific results and enables the creation of new studies based on existing research. The International Union of Pure and Applied Chemistry (IUPAC) defines reproducibility as: "The closeness of agreement between independent results obtained with the same method on identical test material but under different conditions (different operators, different apparatus, different laboratories and/or after different intervals of time). The measure of reproducibility is the standard deviation . . . and a complete statement of reproducibility requires specification of the experimental conditions which differ" (Currie and Svehla 1994). This approach is commonly used to evaluate the reproducibility of experimental studies, and in principle, it is also applicable to evaluating the reproducibility of QSARs. However, this is a complex undertaking and involves QSAR-related research at multiple levels wherein the main criteria to be considered are the reproducibility of the predicted data, the reproducibility of the descriptor data, and the reproducibility of the models.

The IUPAC definition of reproducibility implies some changes in conditions that, ideally, are the presence of different operators in the same or different laboratories over various time intervals. When transferring this definition from experimental methods to QSAR models, it can be concluded that the model is reproducible and truly useful when it can be used independently of the model's authors. It also means that the prediction results are not expected to be identical and that all existing knowledge accounting for the uncertainty of experiments can be applied to QSARs as well. To minimize the uncertainty of model predictions, transparent reporting of QSARs is of critical importance. For this reason, published data on the models should feature technical completeness in reporting so that the most important steps in the model-building process can be independently tested and verified against the reported reference data.

The scientific literature very rarely addresses the issue of the reproduction and reuse of QSAR models that have been published in articles. Dearden et al. (2009) analyzed published models and identified 21 types of errors to be avoided in deriving and presenting published QSAR models. Several of these errors are directly related to the reproducibility of the models, yet no indication is given of how extensive these problems can be. The authors also give numerous recommendations on how to avoid these errors. A relevant study was conducted to verify the reproducibility of the rate constants of hydroxyl radical reaction models by rebuilding them using the same dataset and methodology but with different sets of descriptors (Roy et al. 2011). Recent attempts by the present authors (Ruusmann et al. 2014) to reengineer published QSAR models led to the hypothesis that most results beyond the simplest MLR models are not recoverable and hence are unusable for practical applications. To the best of our knowledge, there are no reported systematic reviews of the documentation, independent verification, and reproducibility of QSAR models presented in the scientific literature. To fill this gap, a review and analysis of scientific articles is needed to assess the completeness of the technical reporting and documentation of QSAR parameters and the representation of the model.

The main aim of this paper is to critically analyze QSAR model publication practices in the scientific literature from the point of view of transparency, potential reproducibility, and independent verification by focusing on the degree of published technical completeness of QSARs. The following section describes a straightforward framework for assessing such data completeness. We then applied this framework to the analysis of a wide variety

of QSAR models that had been developed for various physical and chemical properties, ecotoxicity, environmental fate, human health, and toxicokinetic endpoints. The findings for all of these endpoint groups are summarized and visualized. Based on the analysis, a set of best practices for improving QSAR model reporting and hence the reproducibility of published models is proposed.

## Methods

First, one needs to clarify the terminology. Therefore, within this analysis, a QSAR is considered potentially reproducible when it is transparently published so that the model can be reconstructed and, consequently, independently applied to the recalculation of the QSAR predictions for the chemicals in a dataset with the same or similar accuracy as that in the original paper. Transparency means that the dataset, the computational method description, the model representation, the calculated descriptor values, and the predicted results are explicitly provided with the model. Transparency is important because the availability of data is a prerequisite for the independent assessment of models. However, focusing solely on the transparency of the model's presentation is not sufficient to fully assess whether or not the model is truly reproducible. This study does not attempt a complete independent assessment of the models. Given the volume of literature analyzed, the models were not recalculated; therefore, the reproducibility of the models cannot be fully confirmed. It is therefore correct to say that the model is potentially reproducible. This is a good compromise because testing all the models and data would require too much time and effort. Despite this, during the given work, we independently evaluated a significant number of articles of interest and recalculated their QSAR models (see discussion and examples in Tables 1–5 below), which provided sufficient experience as a basis for the current assessment of the potential reproducibility of published QSARs.

### *Framework of the Literature Analysis*

Among the many online scientific citation indexing services, the subscription-based Web of Science ("All Databases" search over title, abstract, and keywords) was used. The physical and chemical properties, ecotoxicity, environmental fate, human health, and toxicokinetic endpoints listed in the European Commission Joint Research Centre QSAR model database documentation (JRC QSAR Model Database 2017) were analyzed and are presented in the following chapters. Many of these endpoints are required within the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) European Union regulation (EPCEU 2006). For the literature search, endpoint-specific terms were assigned in a systematic fashion and combined with the term "QS*R" (Table S1). The search timespan covered all years up to December 31, 2015. The search results were manually checked to find articles that corresponded to the endpoint of concern and contained original QSAR models (Table S2).

To determine how many articles presenting QSARs contained enough information to verify and potentially reproduce the model (s), the following information was collected in tabular form: *a*) presence of the complete mathematical representation and statistical parameters of the model(s), *b*) presence of chemical structures, *c*) presence of experimental endpoint values, *d*) presence of descriptor values, and *e*) presence of predicted endpoint values. When the article contained the information listed in points *a* through *d* above, it was considered potentially reproducible. The predicted endpoint values (point *e*) could be missing if all the other items were provided because they can be calculated from the given data and used to reproduce the published statistical parameters, whereas their presence gives an extra level of reliability for independent evaluation.

Points *a* to *e* are connected to each other and constitute a comprehensive system. For a reproducible model, it is necessary to have the complete mathematical representation of the model. The presence of chemical structures, which should be presented in the same format as that used for the modeling, is vital for checking that the representation of chemical names and identifiers is correct. The presence of experimental endpoint values, descriptor values, and predicted endpoint values is needed to confirm the validity of the model. The number of articles with complete information on data and models was weighed against the total number of articles with models via calculating percentages for each endpoint and its endpoint group. In addition to the technical completeness of the reviewed QSAR articles, the following details were examined: *a*) the number of articles using more than one mathematical representation for modeling (multiple methods for building models), *b*) the annual distribution of the articles containing models, *c*) the sizes of the datasets used for modeling, and *d*) the prevalence of modeling techniques, wherein all model types were studied.

This framework allowed us to classify the analyzed articles into three groups: not reproducible, potentially reproducible, and a mixture of the two cases. The first group included articles that failed one or more of the above checks. The second group included articles that passed all the checks. The third group included articles containing multiple QSAR models, in which some of the models passed all the checks while others did not.

### *Visualization Methods*

The data collected from the QSAR articles were analyzed, visualized, and discussed with the help of Circos software (version 0.69, by Canada's Michael Smith Genome Sciences Centre, http://circos.ca/software/; Krzywinski et al. 2009), which allows organization and exploration of large and complex datasets and information. Circos plots map data into a circular layout wherein relationships between data elements are highlighted with links, and the link thickness indicates the magnitude of the relationship. The right side of the plot area (see Figure 2 for a legend) is dedicated to highlighting the relationships between the potential reproducibility of articles and the types of modeling algorithms over all years. In all plots (on the right-hand side), the color for YES corresponds to potentially reproducible articles with models, the color for NO corresponds to nonreproducible articles with models, and the color for MIX corresponds to a mixture of both cases. The left side of the plot area is divided into four segments and presents the publishing trends over the past 3 decades. It shows the annual publishing rate of QSAR articles, the technical completeness of the articles, and the distribution of dataset sizes as normalized stacked bar plots. The color coding for technical completeness is the same as that used in the right side of the plot. The color coding for dataset size and total number of such datasets is presented as a legend in the middle of each plot.

## Results

### *Physical and Chemical Properties*

Among the five analyzed endpoint groups, physical and chemical properties were the most widely modeled. This group reported 26 different properties (JRC QSAR Model Database 2017, Table 1), 18 of which were identified in articles containing QSAR models based on a search using a given set of keywords (see Table S1); in total, there were 777 such articles, and they represented >50% of all the articles analyzed. The top five most modeled properties

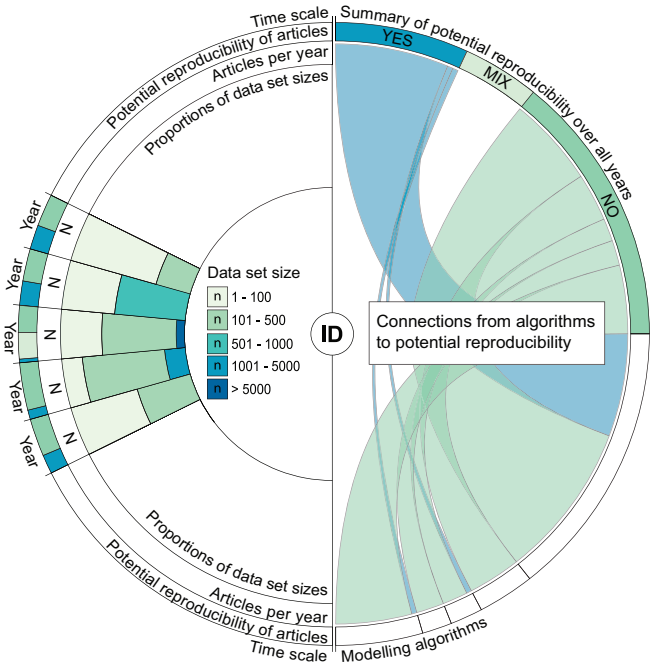


**Figure 2.** Legend for the Circos plot: On the right side of the circle, the color for YES corresponds to potentially reproducible articles, the color for NO corresponds to nonreproducible articles, and the color for MIX corresponds to a mixture of both cases. The left side of the circle is divided into four segments and is dedicated to visualizing the publishing trends over the past 3 decades (first outer-edge strip on the half circle). It shows the technical completeness of articles (second strip on the half-circle) the annual publishing rate of the quantitative and qualitative structure–activity relationship (QSAR) articles (third strip on the half circle), and the distribution of data set sizes as normalized stacked bar plots (fourth strip on the half circle). The color coding for the technical completeness of the articles is the same as that used in the right side of the plot. The color coding for the dataset sizes and the total number of datasets is given as a legend in the middle of the plot.

were boiling point (145), water solubility (131), octanol–water partition coefficient (110), melting point (67), and vapor pressure (65); together, these properties were reported in two-thirds of the articles within the given properties group (see Table 1 for the contributions made by other properties). Other properties were represented in <5% of the identified articles. Preliminary searches for the adsorption/desorption (1.11) property showed results that corresponded with the respective endpoints (2.7 and 2.8) in the environmental fate category (see section "Environmental Fate Endpoints"). Therefore, property 1.11 was not analyzed further because the terms "adsorption" and "desorption" were too general to be used as a search query. Similarly, we did not identify articles reporting complex formation ability in water (1.12), particle size distribution (1.14), fat solubility (1.18), oxidizing properties (1.23), average molecular weight of polymers (1.24), solution/extraction behavior of polymers in water (1.25), and length-weighted geometric mean diameter of fibers (1.26), and therefore did not analyze these endpoints further.

A total of 777 articles were evaluated (see the outer curved strip of the synoptic view in the upper right sector of Figure 3). Of the evaluated articles, 37% (287) qualified as potentially reproducible, i.e., they included enough information that the model could be reused. For an additional 2% (19) of the articles, only the MLR models were reproducible; these were considered partially reproducible. The potential reproducibility of articles within individual properties varied greatly from 21 to 88%. Only three properties had potentially reproducible models in >50% of the published articles (Table 1): for hydrolysis, 88%; for octanol–air partition coefficient, 59%; and for octanol–water partition coefficient, almost 52%. Thus, >60% of the articles (471) in the property group did not include enough information to be considered potentially reproducible.

Most of the articles concerning physical and chemical properties utilized only one modeling algorithm. Exceptions are 92 articles where more than one algorithm was used. MLR models were predominant and were used 653 times (see the outer curved strip in the lower right sector of Figure 3). Among the identified articles that focused on physical and chemical properties as an endpoint, ANNs were used for the first time in 1996 and appeared 139 times within the given timeline. The third most used modeling algorithm was SVM, which appeared in 36 cases. Other machine-learning algorithms in QSARs of physical and chemical properties were used at much lower frequency. For example, in the remaining 70 cases, k-NN (15) and RF (14) algorithms were the most popular. The stacked bar plot at the bottom of Figure 3 shows the use of various modeling algorithms on a timescale. When the number of articles (black line on the stacked bar plot) was compared with the number of modeling algorithms, it was evident that in recent years, more than one algorithm per article was used. In 2003, Dearden (2003) recommended the use of consensus predictions if possible. The number of articles using multiple modeling algorithms has increased since then. In addition, the popularity of studies comparing the predictive capabilities of different algorithms contributes to the increased frequency of the publication of articles using multiple modeling algorithms.

The majority of reproducible models were MLRs (see the interior of the right side of Figure 3, which links modeling algorithms to their potential reproducibility). ANN, the second most common modeling algorithm for physical and chemical properties, was used in 139 models, only 12 of which provided reproducible methods. This low rate was mainly caused by missing reported weights of the neural network neurons. Articles using SVM-derived models shared a similar issue, with only 6 of 36 articles considered potentially reproducible. The use of incompletely reported SVM models is not possible due to the lack of support vector values. When considering the remaining modeling techniques, only 8 of 63 articles were considered potentially reproducible, and the majority of those used polynomial regression models.

The first 15 years (1984–1999) yielded 85 articles, and the models were based on small datasets (<100 data points). Beginning in the year 2000, an average of 43 articles per year were published. The year 2008 marked the largest number of articles per year (63). The datasets used in the modeling (see Table 1 and Figure 3) ranged from very small (5 data points) to very large (58,400 data points). This yielded an average dataset size of 611 compounds, but this value was biased by a couple of very large datasets. The median of 90 compounds reflected the distribution of the dataset sizes more accurately, and this can also be seen in the center of the left-hand side of Figure 3. The distribution of the articles relative to dataset sizes showed that in ~53% of the articles, the dataset size was 100 or smaller; in 30% of the evaluated articles, it was between 101 and 500; in 5% of the articles, it was between 501 and 1,000; in 9% of the articles, it was between 1,001 and 5,000; and in 2% of the articles, it exceeded 5,000 compounds. On an annual basis, in half of the cases, the datasets included 100 compounds or fewer, whereas in a third of the cases, the datasets contained up to 500 compounds, indicating that the QSARs were mostly derived for congeneric datasets. It may be surprising that even for physical and chemical properties, for which experimental data should be easily available, the modeled datasets were small, which suggested that QSAR has been and still is largely mechanistically driven.

**Table 1.** Summary of the literature search for physical and chemical properties.

| Endpoint | | Articles | | | | Dataset size | | | Example(s) |
|---|---|---|---|---|---|---|---|---|---|
| QMRF ID | Name | No. with models | % | No. that are potentially reproducible | Ratio (%) | Min | Max | Median | QDB DOI |
| 1.1. | Melting point | 67 | 8.62 | 13 (1) | 20.90 | 11 | 8,241 | 82 | 10.15152/QDB.127<br>10.15152/QDB.146 |
| 1.2. | Boiling point | 145 | 18.66 | 62 (3) | 44.83 | 10 | 17,768 | 90 | 10.15152/QDB.128<br>10.15152/QDB.122 |
| 1.3. | Water solubility | 131 | 16.86 | 43 (3) | 35.11 | 11 | 42,974 | 136 | 10.15152/QDB.148<br>10.15152/QDB.127<br>10.15152/QDB.173<br>10.15152/QDB.146 |
| 1.4. | Vapor pressure | 65 | 8.37 | 19 | 29.23 | 10 | 1,771 | 107 | 10.15152/QDB.121<br>10.15152/QDB.127<br>10.15152/QDB.173<br>10.15152/QDB.146 |
| 1.5. | Surface tension | 21 | 2.70 | 9 (1) | 47.62 | 10 | 1,604 | 80 | 10.15152/QDB.152<br>10.15152/QDB.122 |
| 1.6. | Octanol–water partition coefficient | 110 | 14.16 | 54 (3) | 51.82 | 8 | 12,831 | 58 | 10.15152/QDB.146 |
| 1.7. | Octanol–water distribution coefficient | 2 | 0.26 | 1 | 50.00 | 24 | 1,130 | 577 | — |
| 1.8. | Octanol–air partition coefficient | 22 | 2.83 | 12 (1) | 59.09 | 7 | 98 | 27 | 10.15152/QDB.146 |
| 1.9. | Air–water partition coefficient | 31 | 3.99 | 9 (1) | 32.26 | 7 | 1,954 | 96 | 10.15152/QDB.150<br>10.15152/QDB.146 |
| 1.10. | Dissociation constant | 22 | 2.83 | 7 (1) | 36.36 | 12 | 58,400 | 61 | 10.15152/QDB.147 |
| 1.11. | Adsorption/desorption | 0 | — | — | — | — | — | — | — |
| 1.12. | Complex formation ability in water | 0 | — | — | — | — | — | — | — |
| 1.13. | Density | 34 | 4.38 | 8 | 23.53 | 5 | 803 | 93 | 10.15152/QDB.122 |
| 1.14. | Particle size distribution | 0 | — | — | — | — | — | — | — |
| 1.15. | Hydrolysis | 17 | 2.19 | 15 | 88.24 | 12 | 40 | 29 | 10.15152/QDB.149 |
| 1.16. | Stability | 14 | 1.80 | 4 (1) | 35.71 | 16 | 99 | 40 | 10.15152/QDB.131 |
| 1.17. | Viscosity | 32 | 4.12 | 9 (2) | 34.38 | 9 | 2,748 | 222 | 10.15152/QDB.151 |
| 1.18. | Fat solubility | 0 | — | — | — | — | — | — | — |
| 1.19. | Flash point | 37 | 4.75 | 14 (1) | 40.54 | 34 | 9,399 | 284 | 10.15152/QDB.123<br>10.15152/QDB.130<br>10.15152/QDB.160 |
| 1.20. | Flammability | 9 | 1.16 | 4 | 44.44 | 543 | 1,615 | 1,038 | 10.15152/QDB.196<br>10.15152/QDB.197 |
| 1.21. | Explosive properties | 9 | 1.16 | 3 | 33.33 | 7 | 227 | 50 | — |
| 1.22. | Autoignition | 9 | 1.16 | 1 (1) | 22.22 | 46 | 820 | 192 | 10.15152/QDB.130 |
| 1.23. | Oxidizing properties | 0 | — | — | — | — | — | — | — |
| 1.24. | Average molecular weight of polymers | 0 | — | — | — | — | — | — | — |
| 1.25. | Solution/extraction behavior of polymers in water | 0 | — | — | — | — | — | — | — |
| 1.26. | Length-weighted geometric mean diameter of fibers | 0 | — | — | — | — | — | — | — |
| | Total | 777 | 100.00 | 287 (19) | 39.38 | — | — | — | — |

Note: Explanation of columns: QMRF ID, quantitative and qualitative structure–activity relationship (QSAR) model reporting format identification number; No. with models, number of articles on physical and chemical properties containing model(s); %, percent of 777 articles; No. that are potentially reproducible, see explanation in the first paragraph of the "Methods" section. The number in parenthesis indicates the number of partially reproducible articles. For ratio (%) these numbers are summed; ratio (%), percent of potentially reproducible articles among the articles with models for a specific endpoint; QDB DOI, digital object identifier in the QSAR Database repository (QSARDB Repository; Ruusmann et al. 2015) for the reproduced article(s). Em dash indicates table cells of endpoints where articles with models were not found or information that can not be provided.

### *Environmental Fate Endpoints*

Although articles with QSAR models were found for all environmental fate endpoints (JRC QSAR Model Database 2017, Table 2), the endpoints were unevenly distributed among the articles. The five most prevalent endpoints accounted for nearly 93% of the 215 reviewed articles; these were bioconcentration (23%), biodegradation (22%), organic carbon–sorption partition coefficient (18%), and abiotic degradation in air (15%) and in water (14%). The results showed where most of the modeling interest has been directed and where enough data for modeling are available. Up to five articles were found for each of the remaining endpoints (Table 2). These scarce endpoints were often modeled as part of experimental measurements.

More than half of the articles (52%) describing environmental fate endpoints contained models that were potentially reproducible (Table 2). In most cases, all models in the article (49%) were transparently presented, while in 3% of the articles, only some of the models were transparently presented (Figure 4, outer curved strip in the upper right sector). Again, the mixed results for potential reproducibility indicate that sufficient information was included for MLRs but not for other types of models. MLR, the most common approach for modeling, was used in 183 articles (Figure 4, curved strip in the lower right sector), of which 109 contained models that were potentially reproducible. The remaining modeling algorithms showed much lower potential reproducibility rates (Figure 4, interior of the right side)—of 22 ANN models, 5 had technically complete documentation, and all 14 SVM models had technically incomplete documentation. In the 37 cases of other modeling algorithms, polynomial regression (8), DTs (7), and k-NN (6) algorithms were more prevalent. Six of these 37 articles were potentially reproducible, and four of the six used polynomial

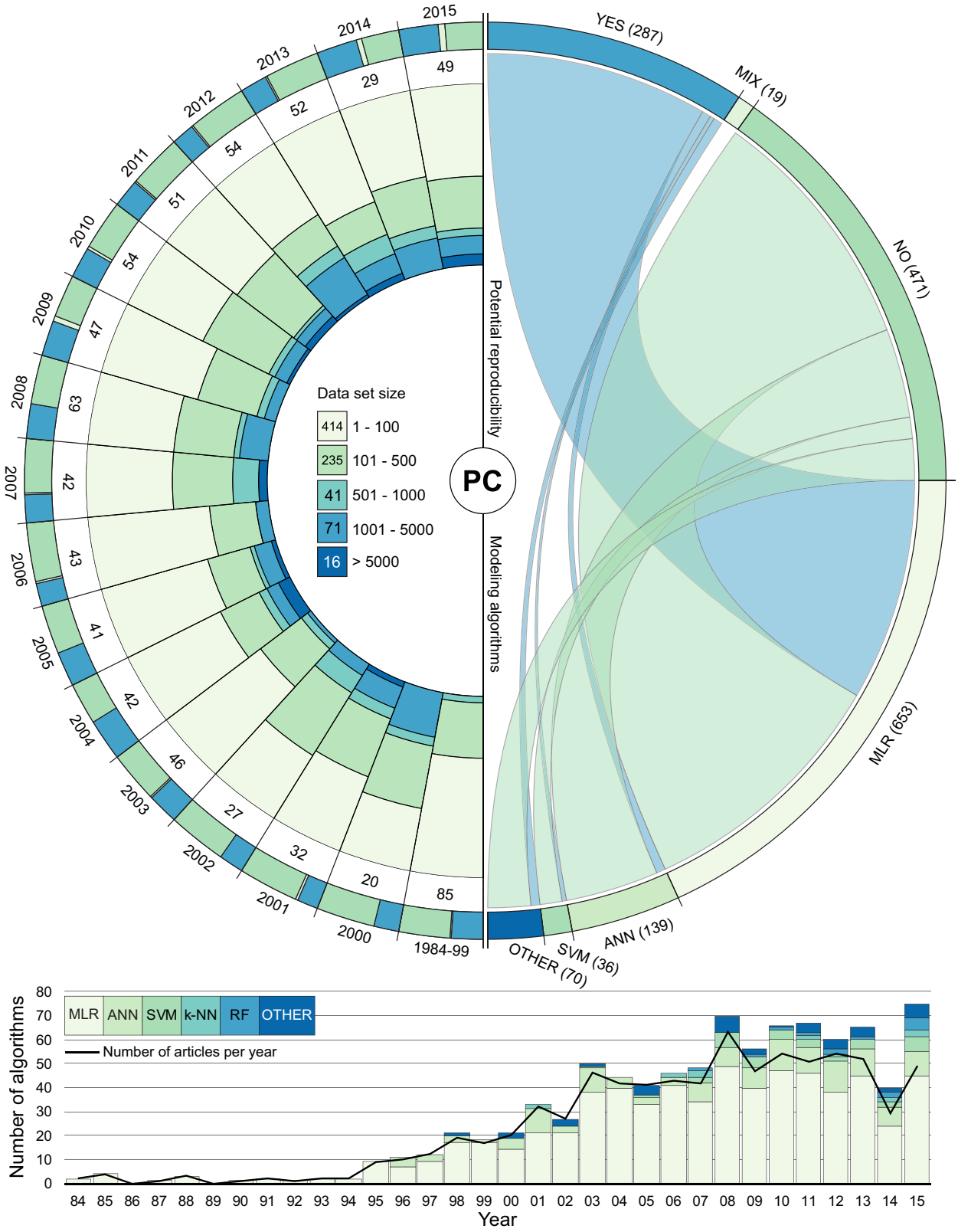


**Figure 3.** Consolidated view of the analysis results for the physical and chemical (PC) properties. Circle: annual distribution of articles (left-hand side), ranges of dataset sizes (in the middle), ratios of potential reproducibility, main modeling algorithms, and relationships between them (right-hand side). Bar plot: annual distribution of articles and modeling methods. Note: ANN, artificial neural networks; k-NN, k-nearest neighbors; MIX, mixture of reproducible and nonreproducible cases; MLR, multilinear regression; NO, nonreproducible articles (models); OTHER, other types of models; RF, random forests; SVM, support vector machines; YES, potentially reproducible articles with models.

**Table 2.** Summary of the literature search for environmental fate endpoints.

| Endpoint | | Articles | | | | Dataset size | | | Example(s) |
|---|---|---|---|---|---|---|---|---|---|
| QMRF ID | Name | No. with models | % | No. that are potentially reproducible | Ratio (%) | Min | Max | Median | QDB DOI |
| 2.1. | Abiotic degradation in water | 30 | 13.95 | 22 | 73.33 | 5 | 1, 431 | 31 | 10.15152/QDB.189 |
| 2.2. | Abiotic degradation in air | 33 | 15.35 | 17 (2) | 57.58 | 7 | 1, 543 | 98 | 10.15152/QDB.201 |
| 2.3. | Biodegradation | 47 | 21.86 | 21 (1) | 46.81 | 7 | 1, 938 | 29 | — |
| 2.4. | Bioconcentration | 50 | 23.26 | 22 (3) | 50.00 | 10 | 1, 036 | 129 | 10.15152/QDB.115<br>10.15152/QDB.110 |
| 2.5. | Bioaccumulation | 3 | 1.40 | 0 | 0.00 | 14 | 49 | 20 | — |
| 2.6. | Organic carbon–sorption partition coefficient | 39 | 18.14 | 16 | 41.03 | 6 | 964 | 66 | 10.15152/QDB.159<br>10.15152/QDB.135 |
| 2.7. | Adsorption/desorption in soil | 5 | 2.33 | 2 | 40.00 | 4 | 53 | 12 | 10.15152/QDB.193 |
| 2.8. | Adsorption/desorption in sediment | 4 | 1.86 | 2 | 50.00 | 8 | 66 | 37 | 10.15152/QDB.194 |
| 2.9. | Vegetation–water partition coefficient | 2 | 0.93 | 2 | 100.00 | 5 | 10 | 8 | 10.15152/QDB.192 |
| 2.10. | Vegetation–air partition coefficient | 1 | 0.46 | 1 | 100.00 | 36 | 36 | 36 | 10.15152/QDB.190 |
| 2.11. | Vegetation–soil partition coefficient | 1 | 0.46 | 1 | 100.00 | 17 | 17 | 17 | — |
| | Total | 215 | 100.00 | 106 (6) | 52.09 | — | — | — | — |

Note: Explanation of columns: QMRF ID, quantitative and qualitative structure–activity relationship (QSAR) model reporting format identification number; No. with models, number of articles with environmental fate endpoints containing model(s); %, percent of 215 articles; No. that are potentially reproducible, see explanation in the first paragraph of the "Methods" section. The number in parenthesis indicates the number of partially reproducible articles. For ratio (%) these numbers are summed; ratio (%), percent of potentially reproducible articles among the articles with models for a specific endpoint; QDB DOI, digital object identifier in the QSAR Database repository (QSARDB Repository; Ruusmann et al. 2015) for the reproduced article(s). Em dash indicates table cells of endpoints where articles with models were not found or information that can not be provided.

regression. Nearly 86% of the articles in this endpoint group used only one modeling algorithm. The timescale for the modeling algorithms and the number of articles (Figure 4, stacked bar plot at the bottom) show that the occurrence of multiple algorithms per article began to increase in 2003 and that it was highest in 2014, when 26% of articles covered more than one modeling algorithm.

The annual distribution of articles on environmental fate endpoints covers a 24-y period from 1991 to 2015 (Figure 4, left side, first curved strip), a much shorter period than the period for articles on the physical and chemical properties. In the 1990s, 25 articles were published (Figure 4, left side, second curved strip). Beginning the year 2000, the publishing rate increased to an average of 12 articles per year. The most active year was 2014, with 23 articles, of which 12 were technically complete. The stacked bar plots (Table 2, Figure 4, left side, interior) reveal that in the years covered by the study, the datasets used in modeling were not large. The smallest dataset contained only four compounds, and the largest included 1,938 compounds. Of the modeled datasets, 60% included <100 compounds (Figure 4, legend in center), 25% of the articles had dataset sizes between 101 and 500, and 12% of the articles had between 501 and 1,000 compounds in the set; only after the year 2010 did the dataset size exceed the threshold of 1,001 data points (this occurred in 3% of the studied articles). An average dataset contained 194 compounds, but again, the real situation is better reflected by the median, which is 53 compounds. Comparison between the number of reproducible articles with models and the dataset sizes of these models revealed that 65% of the articles used datasets that were smaller than median size. Of the articles with potentially reproducible models, 73% used datasets of up to 100 compounds; this increased to 88% when datasets with up to 200 compounds were considered. This indicates that most of the datasets comprised mechanistically similar chemicals and that the models developed for broader ranges of chemicals represent more of an exception.

### *Ecotoxicity Endpoints*

The 261 articles with QSARs for ecotoxicity endpoints (JRC QSAR Model Database 2017, Table 3) were comparable to the number of articles with QSARs for environmental fate endpoints. Of the 13 endpoints, the literature search yielded QSAR articles for 12. QSAR models for toxicity to soil microorganisms (3.7) were not found when the given keywords (Table S1) were used. It was apparent that >82% of the QSAR research effort involving ecotoxicity endpoints has been devoted to the three endpoints of acute toxicity to fish (40%), algae (22%), and *Daphnia* (20%).

Nearly 56% of all ecotoxicity endpoint articles were potentially reproducible, yielding the highest reproducibility rate among the five endpoint groups examined (Table 3). Of the 261 articles, 144 reached technical completeness of data documentation (Figure 5, outer curved strip in the upper right sector), two in the group of partially reproducible articles included sufficient information for MLR models, and 115 lacked sufficient detail to be classified as technically complete. The high verifiability rate of the models in this endpoint group was caused by the small size of the datasets and the simple relationship between toxicity and octanol–water partition coefficient.

The MLR algorithm (221) also dominated the ecotoxicity group (Figure 5, outer curved strip in the lower right sector); it was followed in frequency by ANNs (23) and DT (11). Other modeling algorithms were used in 39 articles; the most popular of these were SVM (8), RF (8), LDA (6), and k-NN (6). Detailed analysis also shows that of the 144 potentially reproducible articles, only three included models that were developed using methods other than MLR (Figure 5, interior of the right side). MLR models were nonreproducible mostly due to missing descriptor values. The timeline of articles and modeling algorithms (Figure 5, at the bottom, stacked bar plot) revealed that multiple algorithms per article were not commonly found for this endpoint group. The most notable exception was the year 2015, when seven articles included up to five different modeling approaches. For comparison, it is worth noting that overall more than one modeling technique was used in 24 articles.

The annual distribution of articles for ecotoxicity endpoints covered the longest period—34 y, ranging from 1981 to 2015 (Figure 5, left side). The proportions of potentially reproducible and potentially nonreproducible articles (the first curved strip) indicate the years in which the reported modeling data were more complete. Before the year 2000, QSAR models were presented in 64 articles (the second curved strip). From 2000 onward, the average publishing rate was 12 articles per year.

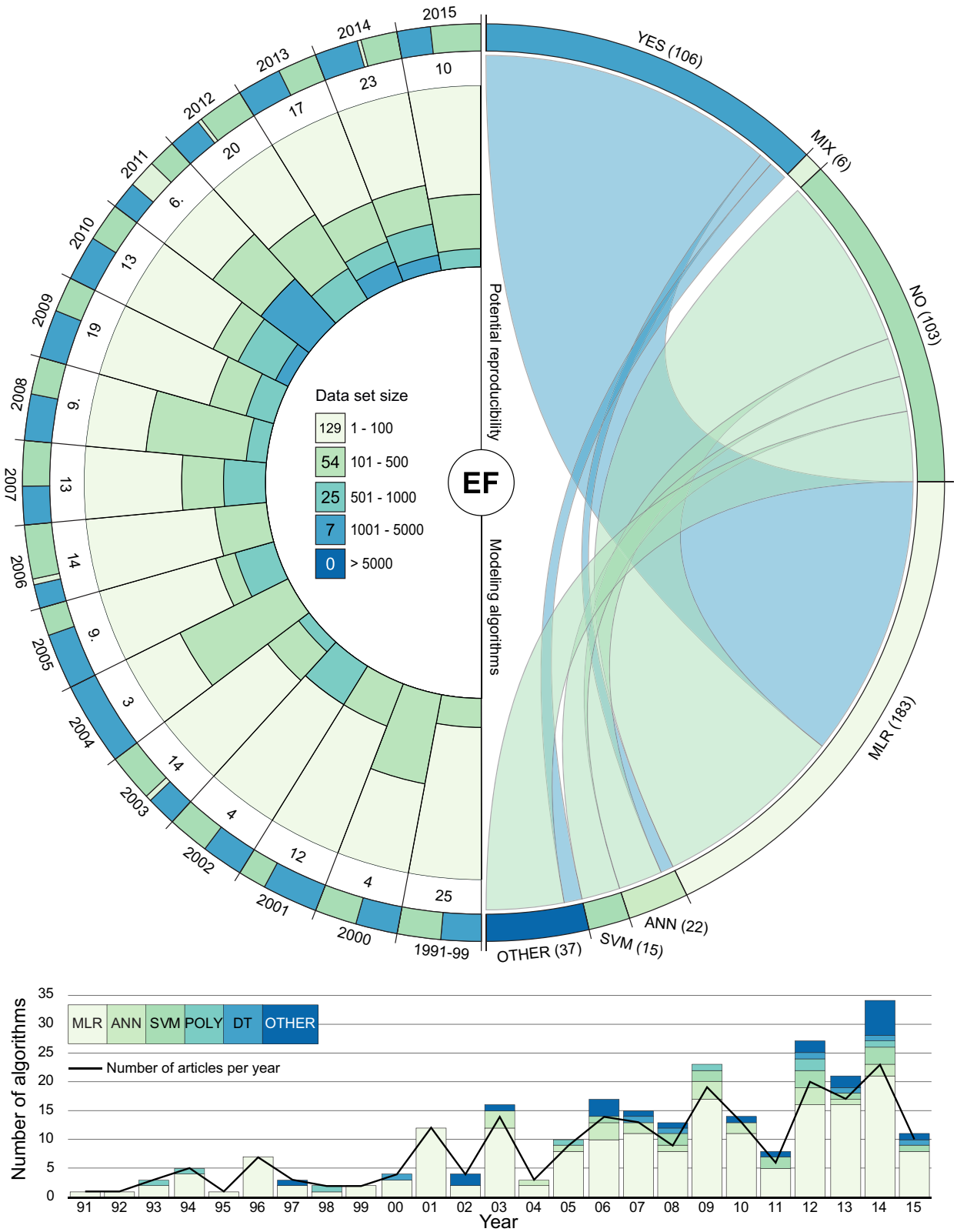


**Figure 4.** Consolidated view of the analysis results for the environmental fate (EF) endpoints. Circle: annual distribution of articles (left-hand side), ranges of dataset sizes (in the middle), ratios of potential reproducibility, main modeling algorithms, and relationships between them (right-hand side). Bar plot: annual distribution of articles and modeling methods. Note: ANN, artificial neural networks; DT, decision trees; MIX, mixture of reproducible and nonreproducible cases; MLR, multilinear regression; NO, nonreproducible articles (models); OTHER, other types of models; POLY, polynomial regression; SVM, support vector machines; YES, potentially reproducible articles with models.

**Table 3.** Summary of the literature search for ecotoxicity endpoints.

| Endpoint | | Articles | | | | Dataset size | | | Example(s) |
|---|---|---|---|---|---|---|---|---|---|
| QMRF ID | Name | No. with models | % | No. that are potentially reproducible | Ratio (%) | Min | Max | Median | QDB DOI |
| 3.1. | Short-term toxicity to Daphnia | 53 | 20.30 | 31 (1) | 60.38 | 9 | 353 | 34 | 10.15152/QDB.111 |
| 3.2. | Short-term toxicity to algae | 58 | 22.22 | 41 (1) | 72.41 | 6 | 873 | 29 | 10.15152/QDB.106<br>10.15152/QDB.195<br>10.15152/QDB.182<br>10.15152/QDB.134<br>10.15152/QDB.144 |
| 3.3. | Acute toxicity to fish | 104 | 39.85 | 43 | 41.35 | 6 | 1,657 | 90 | 10.15152/QDB.195<br>10.15152/QDB.73<br>10.15152/QDB.108 |
| 3.4. | Long-term toxicity to Daphnia | 3 | 1.15 | 2 | 66.67 | 5 | 10 | 10 | — |
| 3.5. | Long-term toxicity to fish | 1 | 0.38 | 1 | 100.00 | 29 | 29 | 29 | 10.15152/QDB.145 |
| 3.6. | Microbial inhibition | 9 | 3.45 | 5 | 55.56 | 8 | 162 | 63 | 10.15152/QDB.200 |
| 3.7. | Toxicity to soil microorganisms | 0 | — | — | — | — | — | — | — |
| 3.8. | Toxicity to earthworms | 2 | 0.77 | 2 | 100.00 | 7 | 11 | 9 | — |
| 3.9. | Toxicity to plants | 12 | 4.60 | 10 | 83.33 | 13 | 42 | 24 | — |
| 3.10. | Toxicity to soil invertebrates | 4 | 1.53 | 3 | 75.00 | 6 | 16 | 8 | — |
| 3.11. | Toxicity to sediment organisms | 5 | 1.92 | 3 | 60.00 | 4 | 12 | 9 | — |
| 3.12. | Toxicity to birds | 4 | 1.53 | 1 | 25.00 | 110 | 663 | 124 | 10.15152/QDB.198 |
| 3.13. | Toxicity to honeybees | 6 | 2.30 | 2 | 33.33 | 45 | 237 | 105 | 10.15152/QDB.157 |
| | Total | 261 | 100.00 | 144 (2) | 55.94 | — | — | — | — |

Note: Explanation of columns: QMRF ID, quantitative and qualitative structure–activity relationship (QSAR) model reporting format identification number; No. with model, number of articles with ecotoxicity endpoints containing model(s); %, percent of 261 articles; No. that are potentially reproducible, see explanation in the first paragraph of the "Methods" section. The number in parenthesis indicates the number of partially reproducible articles. For ratio (%) these numbers are summed; ratio (%), percent of potentially reproducible articles among the articles with models for a specific endpoint; QDB DOI, digital object identifier in the QSAR Database repository (QSARDB Repository; Ruusmann et al. 2015) for the reproduced article(s). Em dash indicates table cells of endpoints where articles with models were not found or information that can not be provided.

The peak number of articles per year (22) occurred in 2006, and the smallest number of articles (4) was published in 2003. Most of the datasets used for modeling were small (see stacked bar plots), and only the last 2 y showed an increase in the number of compounds in the datasets. The smallest dataset contained only four compounds, whereas the largest one included 1,657 compounds and was the only dataset exceeding 1,000 compounds (Table 3). The average dataset size was 132 compounds, and most datasets included <100 compounds. Consequently, the median of 42 compounds better reflects the reality. Figure 5 (left, center of the plot) shows that the dataset size was between 1 and 100 compounds is in 69% of articles; 21% of the articles included between 101 and 500 compounds, and 9% of the articles included between 501 and 1,000 compounds. More than 200 compounds were used in 50 articles. Surprisingly, a dataset as large as 618 for acute toxicity to fish was already available as early as 1991 (Nendza and Russom 1991), while frequent modeling of larger datasets appeared in the early 2000s.

### *Human Health Endpoints*

A literature search for QSARs for human health endpoints (JRC QSAR Model Database 2017, Table 4) identified 142 articles with models; of these, models were found for 16 of a total of 19 endpoints. By far, the most frequently modeled endpoints were carcinogenicity and skin sensitization, with 40 and 28 articles, respectively, containing QSARs; together, these represented ~48% of the articles on human health (Table 4). The executed keywords (Table S1) did not reveal models for photosensitization (4.8), photomutagenicity (4.11), or photocarcinogenicity (4.13).

Of the 142 articles that modeled human health endpoints, 40 articles representing 14 endpoints were deemed potentially reproducible. An additional three articles contained sufficient information to reproduce only the MLR models. Together, these articles represented >30% of all articles with models (see Figure 6, outer curved strip in the upper right sector). The endpoints with the most articles containing sufficient information for potential reproducibility were carcinogenicity (9), skin sensitization (7), and mutagenicity (6). Carcinogenicity and mutagenicity endpoints were modeled in the same articles in several cases due to the close interest in both endpoints.

In this category, 14 articles used more than one modeling algorithm to derive a model. The timescale at the bottom of Figure 6 shows that there was no noticeable increase in the use of multiple algorithms per article through the years studied. MLRs were presented in 75 articles (Figure 6, outer curved strip in lower right sector), ANNs were presented in 21 articles, LDAs were presented in 12 articles, and k-NNs were presented in 9 articles. Compared to the previous endpoint groups in which MLR models were presented in ~85% of the articles, only 52% of the articles this group used MLR models. Notably, human health endpoints included a proportionally larger section of algorithms grouped as "OTHER" compared to the previous endpoints. The most prevalent algorithms in that section were SVMs (7), DTs (7), and logistic regressions (6). Less than half of articles with the MLRs were reproducible (Figure 6, interior of the right side). In addition, there were three LDA models, two DT models, a polynomic model, and a principal component analysis model that could be reproduced. No ANN or SVM model for these endpoints was found to be reproducible because the weights of the neurons or the support vectors, respectively, were not reported in the original article.

The annual distribution of articles on the left-hand side of Figure 6 spans over 24 y, beginning in 1991. An interesting pattern is noted in that until the year 2003, the size of the datasets were small; this was followed by an increase in the size of the datasets until 2015, at which point only one dataset included <100 compounds (Figure 6, stacked bar plot). The difference between the minimum dataset size (4) and the maximum (19,571, endpoint 4.2) was large (Table 4), which is reflected by an average dataset size of 486. Again, the median (69) is a better indicator of the real situation. It is interesting to note that older models (pre-2003) had smaller datasets, but in general, they were easier to reproduce, given that more complete information was presented in those articles. In addition, earlier models

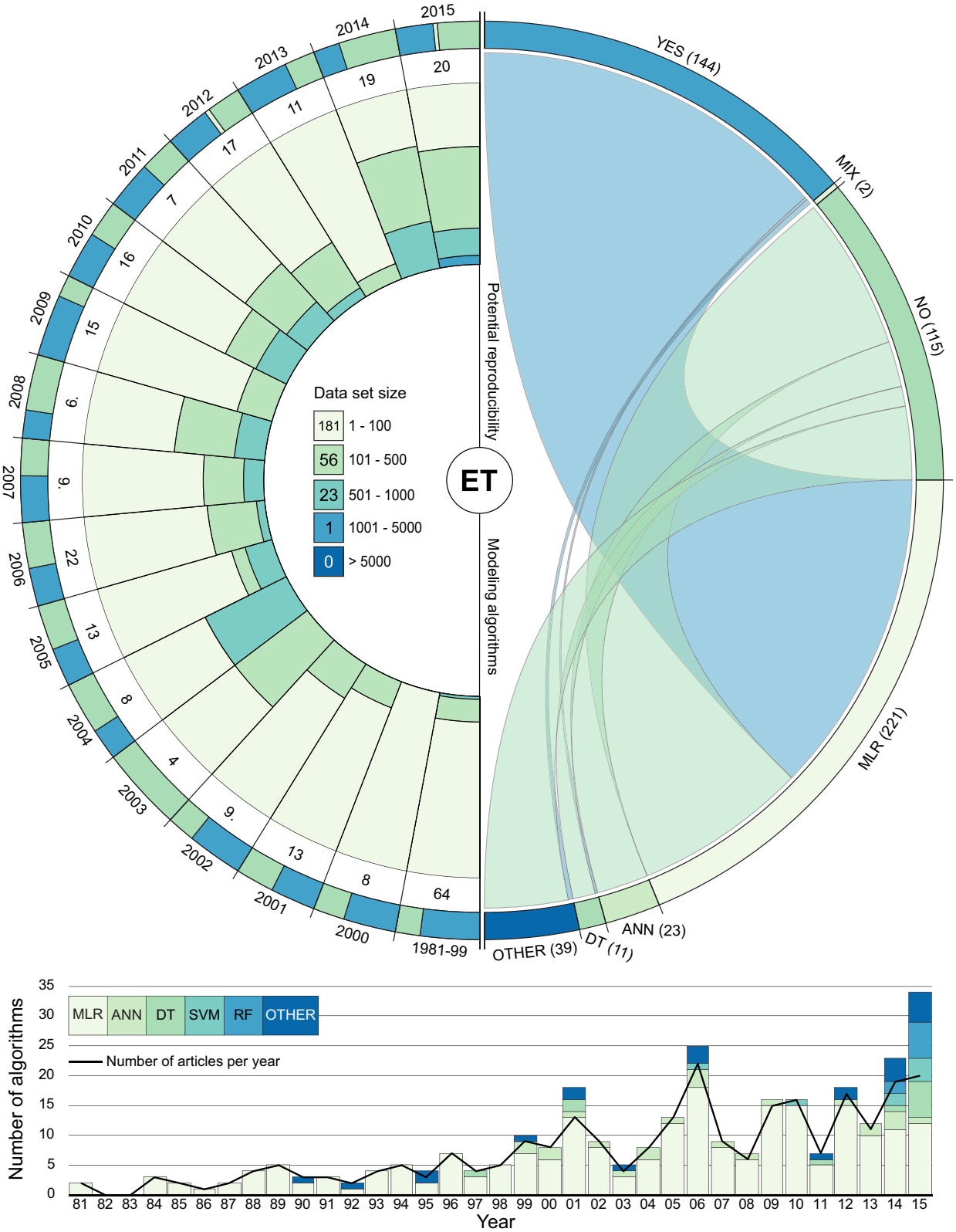


**Figure 5.** Consolidated view of the analysis results for the ecotoxicity (ET) endpoints. Circle: annual distribution of articles (left-hand side), ranges of dataset sizes (in the middle), ratios of potential reproducibility, main modeling algorithms, and relationships between them (right-hand side). Bar plot: annual distribution of articles and modeling methods. Note: ANN, artificial neural networks; DT, decision trees; MIX, mixture of reproducible and nonreproducible cases; MLR, multilinear regression; NO, nonreproducible articles (models); OTHER, other types of models; RF, random forests; SVM, support vector machines; YES, potentially reproducible articles with models.

**Table 4.** Summary of the literature search for human health endpoints.

| Endpoint | | Articles | | | | Dataset size | | | Example(s) |
|---|---|---|---|---|---|---|---|---|---|
| QMRF ID | Name | No. with models | % | No. that are potentially reproducible | Ratio (%) | Min | Max | Median | QDB DOI |
| 4.1. | Acute inhalation toxicity | 3 | 2.11 | 2 (1) | 100.00 | 28 | 108 | 41 | 10.15152/QDB.129<br>10.15152/QDB.179 |
| 4.2. | Acute oral toxicity | 7 | 4.93 | 3 | 42.86 | 27 | 19,571 | 60 | — |
| 4.3. | Acute dermal toxicity | 1 | 0.70 | 1 | 100.00 | 6 | 6 | 6 | — |
| 4.4. | Skin irritation | 5 | 3.52 | 2 | 40.00 | 24 | 2,108 | 186 | 10.15152/QDB.153 |
| 4.5. | Acute photoirritation | 1 | 0.70 | 1 | 100.00 | 53 | 53 | 53 | 10.15152/QDB.139 |
| 4.6. | Skin sensitization | 28 | 19.72 | 7 | 25.00 | 6 | 405 | 139 | 10.15152/QDB.112<br>10.15152/QDB.125 |
| 4.7. | Respiratory sensitization | 4 | 2.82 | 1 | 25.00 | 10 | 319 | 194 | 10.15152/QDB.143 |
| 4.8. | Photosensitization | 0 | — | — | — | — | — | — | — |
| 4.9. | Eye irritation | 15 | 10.56 | 4 (1) | 33.33 | 16 | 2,928 | 52 | 10.15152/QDB.133 |
| 4.10. | Mutagenicity | 14 | 9.86 | 6 | 42.86 | 11 | 6,728 | 159 | — |
| 4.11. | Photomutagenicity | — | — | — | — | — | — | — | — |
| 4.12. | Carcinogenicity | 40 | 28.18 | 8 (1) | 22.50 | 11 | 3,017 | 106 | 10.15152/QDB.142 |
| 4.13. | Photocarcinogenicity | 0 | — | — | — | — | — | — | — |
| 4.14. | Repeated dose toxicity | 6 | 4.23 | 1 | 16.67 | 4 | 549 | 233 | — |
| 4.15. | *In vitro* reproductive toxicity | 1 | 0.70 | 0 | 0.00 | 38 | 38 | 38 | — |
| 4.16. | *In vivo* prenatal developmental toxicity | 3 | 2.11 | 2 | 66.67 | 9 | 39 | 10 | — |
| 4.17. | *In vivo* pre-, peri-, postnatal development and/or fertility | 1 | 0.70 | 0 | 0.00 | 11 | 11 | 11 | — |
| 4.18. | Endocrine activity | 7 | 4.93 | 1 | 14.29 | 19 | 151 | 28 | 10.15152/QDB.124 |
| 4.19. | Neurotoxicity | 6 | 4.23 | 1 | 16.67 | 7 | 58 | 32 | 10.15152/QDB.154 |
| | Total | 142 | 100.00 | 40 (3) | 30.28 | — | — | — | — |

Note: Explanation of columns: QMRF ID, quantitative and qualitative structure–activity relationship (QSAR) model reporting format identification number; No. with model, number of articles with human health endpoints containing model(s); %, percent of 142 articles; No. that are potentially reproducible, see explanation in the first paragraph of the "Methods" section. The number in parenthesis indicates the number of partially reproducible articles. For ratio (%) these numbers are summed; ratio (%), percent of potentially reproducible articles among the articles with models for a specific endpoint; QDB DOI, digital object identifier in the QSAR Database repository (QSARDB Repository; Ruusmann et al. 2015) for the reproduced article(s). Em dash indicates table cells of endpoints where articles with models were not found or information that can not be provided.

tended to show clear relationships between the endpoint and a few descriptors.

### *Toxicokinetics Endpoints*

The toxicokinetics endpoints (JRC QSAR Model Database 2017, Table 5) had the second lowest number of articles (138) with models. QSAR models were found for only six out of 10 endpoints. Gastrointestinal absorption (38% of articles) was the most frequently modeled endpoint in this group, followed by skin penetration (25%), blood–brain barrier penetration (20%), and protein binding (12%). Together, these endpoints represented 95% of the articles in this group. Although articles on DNA binding (5.10) were found, it was difficult to decide which results belonged to this group. Therefore, extraction of relevant articles from the search results was not performed for this endpoint. Unfortunately, for the blood–testis barrier penetration (5.6), blood–lung barrier penetration (5.7), and metabolism (5.8) endpoints, articles with models were not found using the current set of keywords (Table S1).

In this category, 38 articles included the information required to potentially reproduce the model(s), and six articles only included sufficient information to reproduce the MLR models (Figure 7, outer curved strip in upper right sector). Together, these articles comprised nearly 32% of the articles in this group of endpoints. Most of the articles (19 in Table 5) that provided sufficient information to reproduce the models were associated with gastrointestinal absorption (5.3), followed by skin penetration (11), blood–brain barrier penetration (6), protein binding (5), and, finally, ocular membrane penetration (3). Unfortunately, due to their technical incompleteness, neither of the two articles covering placental barrier penetration (5.5) were considered reproducible.

In 30 ($\sim$22%) of the articles in this category, more than one modeling algorithm was used to derive the models. This trend started in 2004 (Figure 7, bottom) and has increased in recent years. The most common modeling algorithm was MLR, with 103 articles (Figure 7, outer curved strip in the lower right sector). ANNs were used in 21 articles, DTs were used in 12 articles, and SVMs were used in 11 articles. Other modeling algorithms [e.g., RFs (8), k-NNs (6), and LDAs (5)] were used 32 times. The relationships between the modeling algorithms and the potential reproducibility of the articles (Figure 7, interior on the right side) again reveal that if the model(s) were potentially reproducible, they were most often derived using MLR. Proportionally, however, this category yielded the second lowest potential reproducibility rate among the evaluated endpoint groups. It is worth mentioning that among DT, LDA, and polynomial regression models, one could observe technical completeness of data, i.e., well-reported models. Unfortunately, full mathematical representation of ANN and SVM models was not present in any of the evaluated articles reporting toxicokinetics endpoints.

This group of endpoints (Figure 7, left side) has been studied with QSARs for the last 20 y. Until 2002, the modeling results presented in articles were well-documented, and the included content was sufficient for reproduction. With the gaining of expertise over the years, one would expect that the quality of model documentation would improve; unfortunately, this is not the case. Over the last 16 y, an average of 8 articles per year were published (second curved strip). In the most productive year, 2007, in which 15 articles were published, there was surprisingly low technical completeness of the reported model data—only one article was potentially reproducible, and two articles comprised mixtures of potentially reproducible and nonreproducible models. The largest number of potentially reproducible articles was published in 2002; five of the seven articles published in that year included complete technical data. The datasets used in modeling varied in size (see Table 5 and Figure 7, stacked bar plots)—the smallest dataset contained five compounds, and the largest had 20,795 compounds. The average number of compounds in the datasets was 335. The median number of compounds in the

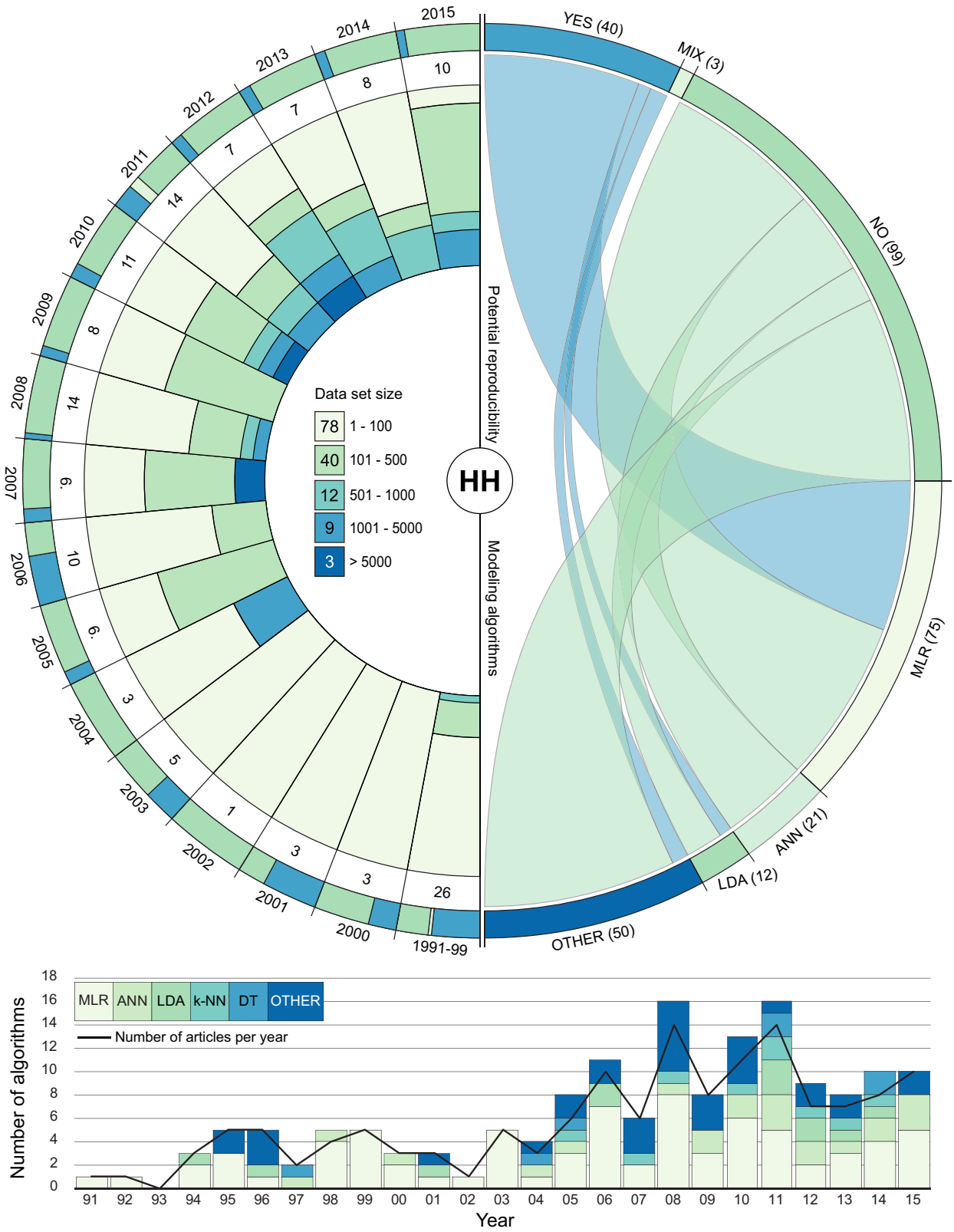


**Figure 6.** Consolidated view of the analysis results for the human health (HH) endpoints. Circle: annual distribution of articles (left-hand side), ranges of data-set sizes (middle section), ratios of potential reproducibility, main modeling algorithms, and relationships between them (right-hand side). Bar plot: annual distribution of articles and modeling methods. Note: ANN, artificial neural networks; DT, decision trees; k-NN, k-nearest neighbors; LDA, linear discriminant analysis; MIX, mixture of reproducible and nonreproducible cases; MLR, multilinear regression; NO, nonreproducible articles (models); OTHER, other types of models; YES, potentially reproducible articles with models.

**Table 5.** Summary of the literature search for toxicokinetics endpoints.

| Endpoint | | Articles | | | | Dataset size | | | Example(s) |
|---|---|---|---|---|---|---|---|---|---|
| QMRF ID | Name | No. with models | % | No. that are potentially reproducible | Ratio (%) | Min | Max | Median | QDB DOI |
| 5.1. | Skin penetration | 35 | 25.36 | 9 (2) | 31.43 | 5 | 454 | 111 | — |
| 5.2. | Ocular membrane penetration | 4 | 2.90 | 3 | 75.00 | 9 | 69 | 47 | 10.15152/QDB.191 |
| 5.3. | Gastrointestinal absorption | 53 | 38.41 | 17 (2) | 35.85 | 17 | 1,301 | 100 | 10.15152/QDB.166 |
| 5.4. | Blood–brain barrier penetration | 28 | 20.29 | 5 (1) | 21.43 | 18 | 484 | 143 | 10.15152/QDB.199 |
| 5.5. | Placental barrier penetration | 2 | 1.45 | 0 | 0 | 88 | 88 | 88 | — |
| 5.6. | Blood–testis barrier penetration | 0 | — | — | — | — | — | — | — |
| 5.7. | Blood–lung barrier penetration | 0 | — | — | — | — | — | — | — |
| 5.8. | Metabolism | 0 | — | — | — | — | — | — | — |
| 5.9. | Protein binding | 16 | 11.59 | 4 (1) | 31.25 | 10 | 20,795 | 144 | — |
| 5.10. | DNA-binding | 0 | — | — | — | — | — | — | — |
| | Total | 138 | 100.00 | 38 (6) | 31.88 | — | — | — | — |

Note: Explanation of columns: QMRF ID, quantitative and qualitative structure–activity relationship (QSAR) model reporting format identification number; No. with model, number of articles with toxicokinetics endpoints containing model(s); %, percent of 138 articles; No. that are potentially reproducible, see explanation in the first paragraph of the "Methods" section. The number in parenthesis indicates the number of partially reproducible articles. For ratio (%) these numbers are summed; ratio (%), percent of potentially reproducible articles among the articles with models for a specific endpoint; QDB DOI, digital object identifier in the QSAR Database repository (QSARDB Repository; Ruusmann et al. 2015) for the reproduced article(s). Em dash indicates table cells of endpoints where articles with models were not found or information that can not be provided.

datasets was 103, indicating that most of the modeling utilized small data collections. This can be clearly seen in Figure 7 (center of the plot), where the dataset size was <100 compounds in 50% of the cases and between 101 and 500 compounds in 41% of the cases. Datasets with higher numbers of compounds were an exception. Datasets with 501 to 1,000 compounds were found in eight articles; between 1,001 and 5,000 compounds were used in three articles, and >5,000 compounds were used only once. The size of the datasets has clearly increased over the past decade.

## Discussion

Depending on the availability and transparency of the data and model information, QSAR articles can be divided into three groups: not reproducible, potentially reproducible, and a mixture of the two cases. The first group, articles with models that cannot be reproduced, was the easiest to detect because those articles lacked some critical data. Reproducibility can fail due to incomplete or even completely missing representation of the chemicals in the dataset (e.g., identifiers of chemical structure, descriptors, and/or experimental and calculated endpoint data). For example, without descriptor values, it is impossible to confirm the reproducibility of the descriptor calculation procedure, the correctness of the model representation, or the predicted endpoint values. Finally, a model without a complete mathematical representation cannot be used, making it impossible to confirm the reproducibility of the predicted values and the model's statistical parameters. Unfortunately, such articles are in the majority, and the "Results" section vividly shows this. It can be argued that some models with missing information might be reproducible with hard work by following the protocol described in the scientific article. This effort may include collecting data from the original sources and recalculating everything from the possible conformational search of the chemical structures to the generation of the QSAR model using the original software (and version) used by the authors of the article. This approach might work with acceptable accuracy for simple models (e.g., MLR models for small datasets), but it is very unlikely to be successful for large datasets or for models that use stochastic methods.

The second group includes potentially reproducible articles. This is the most useful outcome of a scientific article and should be the norm. Such articles report all technical and numerical data corresponding to each step of the model's derivation and are ready for independent verification. Finally, if everything is correct, these reported models are beneficial to the users. In such a case, the user can reproduce the model, use the same version of the software or an alternative software package as a descriptor calculator, or create a workflow and execute the model for new compounds that need prediction. The results of this work show proportionally how many of such studies are reported in the scientific literature and how well the articles document them.

The third group includes articles that represent a mixture of the first and second cases. Such articles usually compare different modeling methods and contain multiple models for which the representation of one or more models has technical completeness (e.g., MLR), but the remaining models may have incomplete or missing mathematical representations. Most often, in these cases, the authors have probably not been able to find a good way to present or document models that exploit more complex machine-learning algorithms. Perhaps there are limitations to the presentation of mathematical or chemical data in journals or limitations may be imposed by the authors' lack of expertise or sometimes by a lack of full reporting documentation on the software employed by the authors (the black box problem).

### *Observations across the Endpoints*

The systematic analysis of 1,533 published scientific articles involving QSAR models covering various physical and chemical properties, ecotoxicity, environmental fate, human health, and toxicokinetics endpoints brings forward a worrying trend—overall, 57.5% of QSAR articles lacked some details about the dataset or the model representation that would make the independent verification and reproduction of these models impossible or very difficult. Thus, only 42.5% of the articles presented potentially reproducible models. One should keep in mind that this is an optimistic estimate because it assumes that all of these models can actually be reproduced. Our experience in rebuilding previously published QSARs shows that for several unforeseen reasons, this is unlikely to be true and that the percentage of reproducible models is actually even lower (see discussion below).

Estimating the exact number of articles with QSAR models for each endpoint was not the main goal of the literature search methodology used in this work. The set of keywords chosen for the literature search influenced the results. Sometimes, even the use of the most clearly understood keywords does not ensure that all existing articles will be identified because search engines have limitations. A good example is our recently published article on a QSAR for bioconcentration factor (Piir et al. 2014); that article did not appear in the search results because the term "QS*R" was neither present in the abstract nor specified as a keyword. For that reason, that article and probably several other articles were

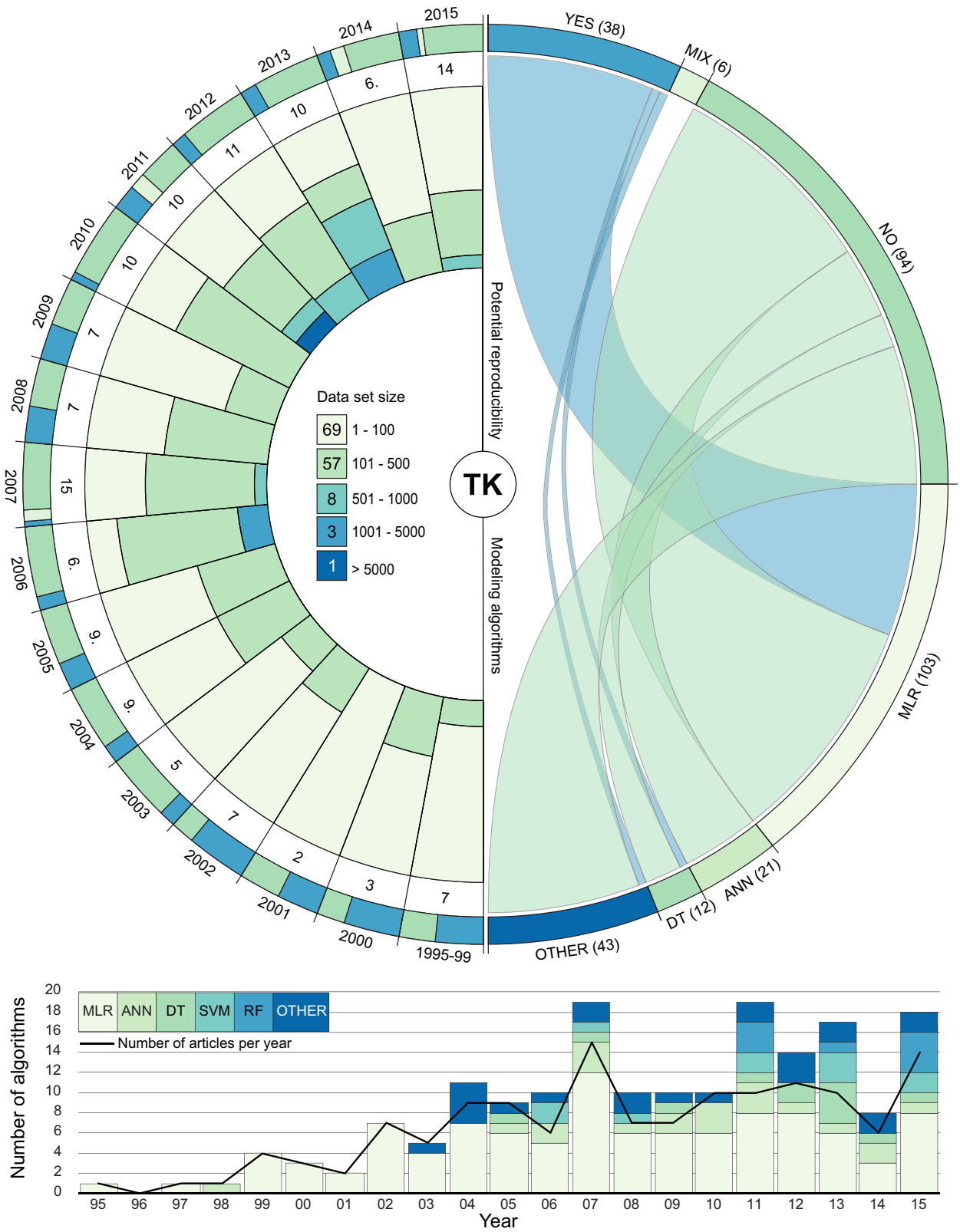


**Figure 7.** Consolidated view of the analysis results for the toxicokinetics (TK) endpoints. Circle: annual distribution of articles (left-hand side), ranges of data-set sizes (in the middle), ratios of potential reproducibility, main modeling algorithms, and relationships between them (right-hand side). Bar plot: annual distribution of articles and modeling methods. Note: ANN, artificial neural networks; DT, decision trees; MIX, mixture of reproducible and nonreproducible cases; MLR, multilinear regression; NO, nonreproducible articles (models); OTHER, other types of models; RF, random forests; SVM, support vector machines; YES, potentially reproducible articles with model.

not included in the analysis. Hence, the visibility and discoverability of published QSARs in search engines is also determined by the pertinence and consistency of words used in the keywords, titles, and abstracts of these articles. A systematic review might have also included more alternative keywords in the search (e.g., "structure–activity relationship" or "relationship," etc.) to make it possible to find all relevant articles in the literature, but this is not the scope of this manuscript. In the present work, we wanted to design a minimal set of keywords to reduce false hits and ensure that the literature search could be easily repeated and verified. Nevertheless, the number of reviewed and analyzed articles (1,533 of the 6,254 search hits) was feasible for a statistical analysis and allowed us to determine how well QSAR models are documented in scientific articles, as well as the proportion of QSARs that can be potentially reproduced and reused.

Here, a minority of endpoints in each endpoint category received the most attention. This was obviously influenced by the availability and quality of experimental data. Analysis of the dataset sizes showed that data availability is a general issue, particularly for the human health endpoints. The small number of papers per year in the human health endpoint group is most likely related to the lack of available data. This stands in contrast to the strong interest in models of human health endpoints given their relevance to toxicity in humans. The paucity of data may be related to the ethical implications of the use of higher living organisms, such as rats and dogs, as well as to the high cost of complex tests and the regulations regarding such tests. Another possible reason for the paucity of publicly available data is the push toward the commercialization of toxicity-related models.

In this analysis, MLR models predominated (Figure 8) and also had the highest potential for reproducibility compared to other mathematical representations. This does not mean that the situation concerning MLR models is satisfactory; in fact, only 51% were presented in a technically complete manner. For all other model types, the potential for reproducibility was found in only 10% of the published articles. This does not support our earlier hypothesis that complex models are not, as a rule, reusable on the basis of the information presented in the article (Ruusmann et al. 2014). In fact, we identified several articles with more complex model types (ANNs, SVM-s, k-NNs, DTs, RFs, etc.) that included complete mathematical representations, but they were in the minority.

The present work suggests several reasons why many models cannot be independently verified. For example, in the case of MLR models, the mathematical representation was usually given correctly, yet problems can be caused by missing descriptor and/or property values. In the case of more complex models, the mathematical representation was often missing, as were the descriptor values. Indeed, the representation of complex models can be complicated, but there are solutions that can make it easy to represent models transparently and make them usable. To encourage the publication of easily accessible models, examples of MLRs as well as of more complex models for the majority (54 of 79) of the analyzed endpoints have been reproduced from scientific articles and are accessible through the QSAR DataBank Repository (Ruusmann et al. 2015; QSARDB Repository); this information is archived in the QSAR DataBank file format (Ruusmann et al. 2014). See the respective DOIs of the QDB archives in Tables 1–5.

Large data tables take up space; therefore, many authors have made compromises when providing them in articles, often omitting descriptor values. As a consequence, more complete technical data are available for QSARs that contain smaller numbers of compounds and fewer descriptors. This is an understandable outcome, given that it is easy for authors to provide full data for such models in the article. The situation has improved during the last decade with the increased availability of online supplementary information, allowing complete data tables for large datasets to be published. The results of this study suggest that, unfortunately, even in a time with increasing acceptance and use of supplementary material to report data, there is still a lack of reproducibility among QSAR models due to incomplete documentation. While supplementary information is highly useful, it has a number of technical limitations, depending on the journal. For example, some publishers accept only a limited set of file formats, impose file size restrictions, and so on. To address the above issues, open data repositories are a good alternative for long-term preservation of primary chemistry research data (Downing et al. 2008). In recent years, the open repository movement has created new tools for making data and models available in domain-specific, content-aware smart repository solutions designed for storing QSAR models (Ruusmann et al. 2015).

Even if the data appear to be well presented, independent evaluation may uncover a nonobvious inconsistency in the data that will make a model unusable. The following warning applies mostly to authors, but reviewers in the peer-review process may benefit as well. Several of our attempts to reproduce models from articles have failed due to inconsistent data and/or incorrect mathematical representations. Similar deficiencies have also been noted by Dearden et al. (2009). Those mismatches were probably caused by copy-and-paste errors. Therefore, authors must cross-check their tables. Detecting such errors is quite easy in the case of MLR models. It only requires recalculation of the predicted values and checking whether they match the provided values. Further, omitting information about data preprocessing (normalization, etc.) from articles makes it difficult to properly recreate the proposed model.

In 2006, the *Journal of Chemical Information and Modeling*, together with the *Journal of Medicinal Chemistry*, revised their publication policy for QSAR manuscripts. Along with several policy changes intended to improve the quality of QSAR papers, the editors have stated the following fundamental requirement for manuscripts (Jorgensen 2006; Editorial 2006): "All data and molecular structures used to carry out a QSAR/QSPR study are to be reported in the paper and/or in its Supporting Information, or be readily available, without infringements or restrictions. The use of proprietary data is generally not acceptable because it is inconsistent with the ACS Ethical Guidelines for publications: 'A primary research report should contain sufficient detail and reference to public sources of information to permit the author's peers to repeat the work.' This is fundamental, though possible exceptions can be discussed with the editor in the unusual circumstance that a convincing case could be made that the data are somehow a secondary issue." In fact, similar principles for the reproducibility of results are universally required by most journals. Unfortunately, the results of this study demonstrate that this fundamental requirement is often still overlooked or has not been sufficient.

QSAR modeling approach has been accepted for a long time in industry, e.g., for drug design. It can be expected that in the future, the use of QSARs by regulators to estimate the ecological and human health effects of chemical substances will increase. For example, the REACH (EPCEU 2006) and Cosmetics (EPCEU 2009) regulations imposed by the European Union promote the use of *in silico* methods to reduce animal testing. The general driver in the use of QSARs is, however, chemical safety, which is the main reason for the attrition of chemicals from the market. The acceptance and use of QSARs by regulators and various industries will require thorough validation and assessment of models (OECD 2007). This cannot be achieved without transparent and accurate

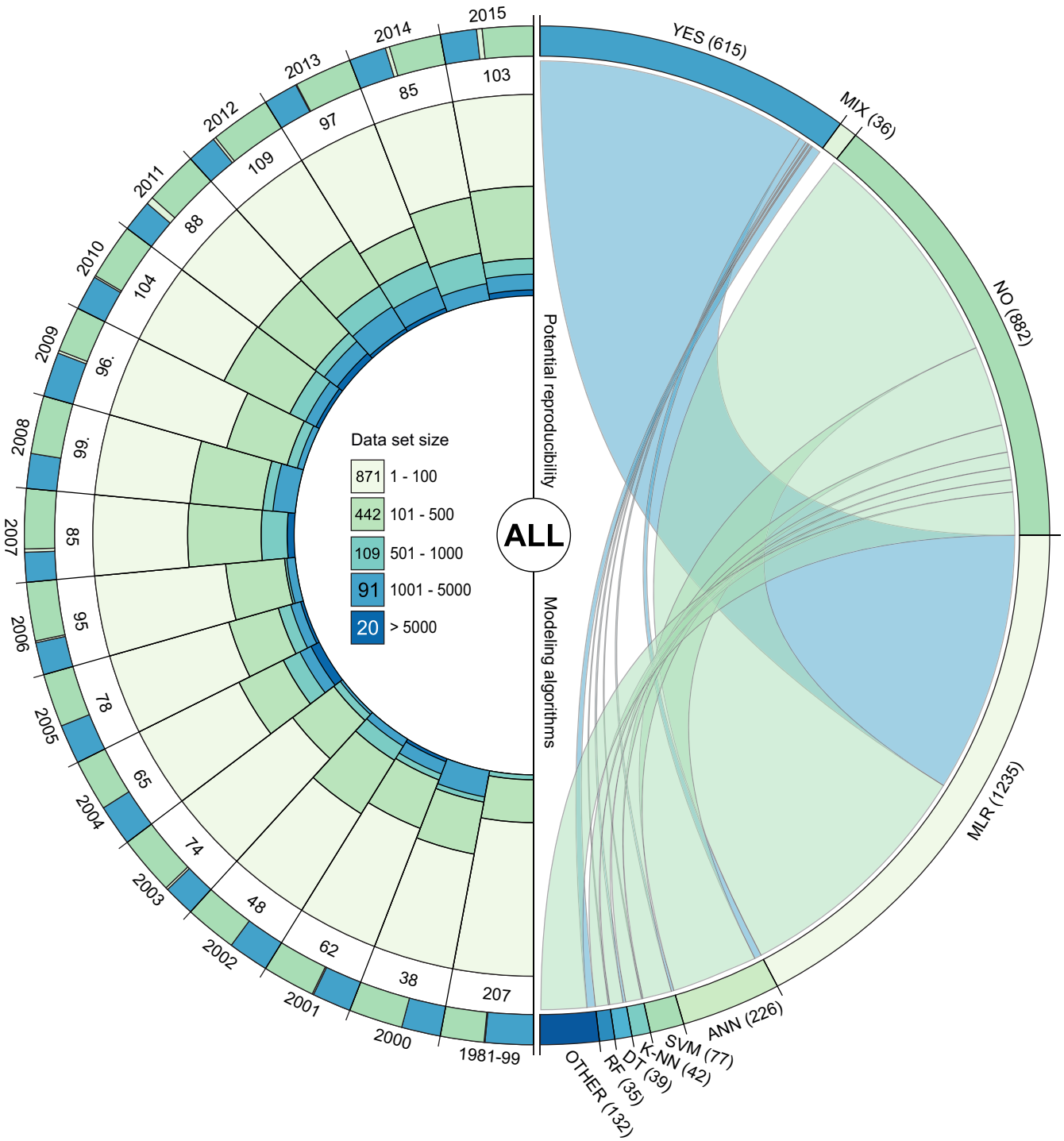


**Figure 8.** Consolidated view of the literature review and analysis results. Circle: annual distribution of articles (left-hand side), ranges of dataset sizes (in the middle), ratios of potential reproducibility, main modeling algorithms, and relationships between them (right-hand side). Note: ANN, artificial neural networks; DT, decision trees; k-NN, k-nearest neighbors; MIX, mixture of reproducible and nonreproducible cases; MLR, multilinear regression; NO, nonreproducible articles (models); OTHER, other types of models; RF, random forests; SVM, support vector machines; YES, potentially reproducible articles with models.

modeling and model reporting practices. In the case of decision-making, transparency enables reasoning on why the prediction is applicable to a test compound, and independent validation gives practical value to the model. This makes the best practices of model reporting (see next section) a basic requirement when publishing scientific work involving QSAR models. The criteria used in the present review for the assessment of scientific articles for potential reproducibility correlate very well with OECD guidance on the validation of QSAR models (OECD 2007). Directly speaking, if a research article is reproducible and includes a proper mechanistic interpretation of the published model(s), it will very likely meet all of the OECD QSAR validation principles. This is true for several reasons, with the main being that the model can be independently verified, including all calculation procedures and the applicability domain. Based on the present review, one can say that up to 42.5% of published QSAR models potentially comply with the OECD

QSAR validation principle. However, this is an overly optimistic estimate because each individual case would require time-consuming independent verification, and verification is not likely to succeed for every model.

Fortunately, some QSAR models have found their way into practical application in commercial or public software solutions, where all molecular structure and data transformations used in the model development workflows are also realized in the prediction functionality of these programs (Nicolotti et al. 2014; Ambure et al. 2014; Tetko et al. 2017). In principle, this should guarantee the reproducibility of these models and make them easily usable. An obvious drawback is that development, maintenance, and support of the software are not trivial, and each of these activities involves costs that are not simple to sustain for a large number of models. As a result, the majority of QSAR models are not available as ready-to-use software, while at the same time, they have value that can be realized in the hands of an expert user. To make published QSAR models more easily accessible than they have so far been, better documentation of these models is the key.

### *Best Practices for QSAR Model Reporting*

The current analysis clearly indicates substantial deficiencies in the documenting of QSAR studies in scientific articles. The incidence of technical completeness of reporting and the reproducibility of derived models is less than half in all cases (Figure 8, right-hand side). This suggests that in addition to best practices on how to develop the models (see "Introduction" section for relevant references), best practices on how to report the models are needed.

Therefore, based on our prior experience in this field, the results of the literature analysis, and experience gained by reproducing models from the scientific literature (examples are provided in Tables 1–5), a checklist for assessing the potential reproducibility of published models was compiled (Figure 9). This checklist does not aim to capture every possible detail related to the description of the model. Instead, it captures the minimal requirements needed to make the model reproduction process possible and the model potentially usable by a wider interested audience. The checklist can also be used by interested parties, such as journal editors, reviewers, regulators, evaluators, and potential users, to assess the reproducibility of articles with models. It is focused on raw data, data provenance, and model representation. By following this simple list, it is easy to improve the transparency of published QSAR research, allowing broader applicability of the research results and fulfilling the basic requirements of reproducibility. The checklist is divided into four parts according to the major subjects that are typical of a QSAR development workflow (Sild et al. 2006; Maran et al. 2007; Tropsha 2010). All parts of the checklist are equally important, and a deficiency in any of them will significantly compromise the reproducibility (i.e., independent verification) and, as a consequence, will compromise the use of the models.

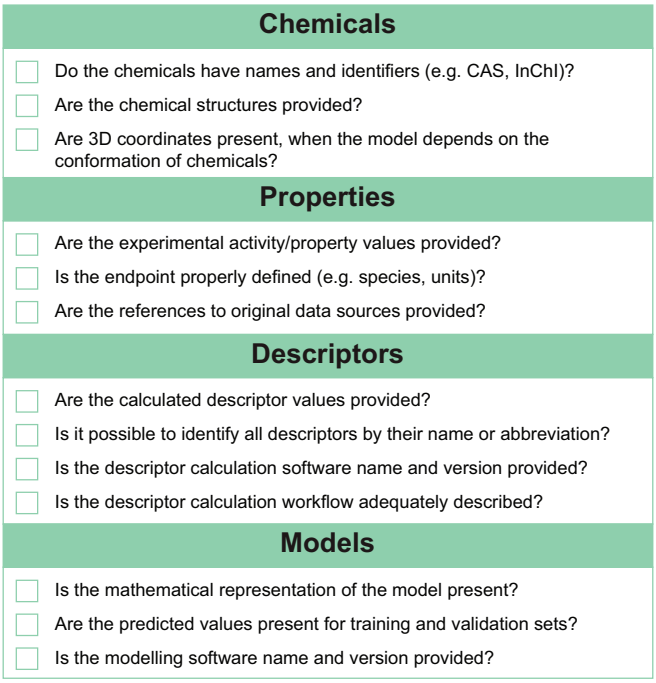


**Figure 9.** Checklist for the articles of quantitative and qualitative structure–activity relationships (QSAR) models.

The first part of the checklist (Figure 9) covers the characterization of chemicals and their molecular structures used in model development. This is a two-sided problem because it involves identification and characterization of the chemicals. Correct identification enables the mapping of chemical structures to experimental measurements and requires that chemical names (ideally, supplier provided), Chemical Abstract Service (CAS) registry numbers, International Chemical Identifier (InChI) codes, or other identifiers are correct and consistent with each other (see discussion in Ruusmann and Maran 2013). The proper characterization of chemicals can be a major concern for reproducibility because it is not always obvious how to determine the molecular structure of a compound from its chemical identity. Results can vary since different laboratories utilize different protocols to process chemical structural data. Discrepancies in the results may arise from different ways of handling salts, mixtures, tautomers, and stereoisomers, generating 3-D coordinates with stochastic algorithms, and so on. Therefore, it is strongly recommended to make all manipulations conducted during the structure pretreatment explicitly available, together with chemical structures in the file format used during modeling.

The second part covers the representation of experimental data. From the point of view of reproducibility, the most important question is whether or not the experimental values are presented. However, the requirement does not end there; it is also important to define the measured property (or endpoint) and identify the original data sources. Information on whether the data are native or synthetic (see discussion in Ruusmann and Maran 2013) must be presented, and in the latter case, all manipulations must be described. When publishing experimental data, it is advisable to avoid PDF files and give preference to machine-readable file formats (e.g., text files such as comma/tab–separated values and spreadsheets).

The third part covers the descriptors that are used in the model. The availability of descriptor values is critical to verifying the descriptor calculation procedure. In addition, the identifiers and names of the descriptors in the article must be consistent with the software that was used. It is extremely important to identify the version of the software that was used because the descriptor values may change due to bug fixes and improvements to the underlying algorithms. Similarly to the data on experimental properties, it is important to consider the choice of file formats to be used in publishing the descriptor values.

The fourth part covers the representation of the QSAR models and the procedure used for model development. The mathematical representation of the model is essential for performing predictions. However, the current literature analysis shows that this requirement is usually only satisfied for MLR models because they can easily be added to the article text as mathematical equations. The description of the model development process must contain the name and version of the modeling software, together

with the parameter values specific to the modeling technique used to build the model. In addition, the availability of the predicted property values is important because they enable the verification of each prediction when model reproduction is attempted.

When carefully documented, all of the above described information can be provided in the scientific article. For smaller data series and simpler mathematical representations (like MLR), this is easily doable. In the case of larger datasets and more complex mathematical models (such as machine-learning algorithms), including all relevant information in the article will become more complicated, and with it, data presentation becomes a challenge. Presentation of full data in the article text may not be very practical, but the presentation of such data is easily achievable when the data are provided in the supplementary material in a format that is specifically designed to represent the models. A recent book chapter provides an overview of data formats that have been developed for organizing QSAR models and related information (Sild et al. in press). The QsarDB data format fulfills all of the above requirements for the representation of QSAR models and datasets and is probably the most complete data format developed to date for QSAR models (Ruusmann et al. 2014; QSARDB Repository, see Best Practises menu item under the Resources). Most importantly, the format is machine readable and enables independent evaluation and use of the models to make predictions when descriptor values are provided, even in the case of machine-learning methods that are not easily applicable to prediction based on the description of the model in a scientific article. The QsarDB data format is a collection of systematically organized text files in well-known file formats for representing chemical structures, tabular data, literature references, and so on, including the model itself in an executable format. The QsarDB data format is supported by a graphical user editor and tools that can be used to organize all pertinent information into catalogs and, finally, into a zip archive.

## Conclusions

This review highlights critical issues in the scientific literature regarding QSAR parameters and models. The results demonstrate that the fundamental requirements established by many leading scientific journals $\sim$10 y ago for reproducible documentation of QSAR papers are often still being overlooked or have not been sufficient. Hopefully, the results of this investigation and our recommendations regarding best practices for model reporting will improve the reproducibility and usefulness of future QSAR studies for interested parties (journal editors, reviewers, regulators, evaluators, and potential users).

## Acknowledgments

Estonian Ministry of Education and Research [grant IUT34-14]. European Union European Regional Development Fund [Foundation Archimedes (E.E.)] [grant number 3.2.1201.13-0021; grant number TK143 (Centre of Excellence in Molecular Cell Engineering)].

G.P. performed the literature search on physical and chemical properties and toxicokinetics endpoints, wrote text for the respective parts of the manuscript, double-checked other endpoints, performed the final data analysis, prepared Circos plots, adapted other illustrations, and contributed to other parts of the text. I.K. performed the literature search on environmental fate and ecotoxicity endpoints and contributed to the fine-tuning of the text throughout the manuscript. A.T.G.S. performed the literature search on human health endpoints, analyzed data, wrote text for the respective part of the manuscript, and contributed to other parts of the text. S.S. worked on text for the reproducibility of models and best practices of QSAR model reporting and contributed to other parts of the text. P.A. performed the literature search on some environmental fate and ecotoxicity endpoints. U.M. conceived the review, wrote the introduction and parts of the discussion in several places and contributed to the text throughout the article. All authors participated in discussions during the preparation of the manuscript. All authors have read and approved the final manuscript.

## References

Alexander DLJ, Tropsha A, Winkler DA. 2015. Beware of R2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. J Chem Inf Model 55(7):1316–1322, PMID: 26099013, https://doi.org/10.1021/acs.jcim.5b00206.

Ambure P, Aher RB, Roy K. 2014. Recent advances in the open access cheminformatics toolkits, software tools, workflow environments, and databases. In: *Computer-Aided Drug Discovery. Methods in Pharmacology and Toxicology*. Zhang W, ed. New York, NY:Humana Press, 257–296, https://doi.org/10.1007/7653_2014_35.

Benfenati E, Clook M, Fryday S, Hart A. 2007. QSARs for regulatory purposes: the case for pesticide authorization. In: *Quantitative Structure-Activity Relationship (QSAR) for Pesticide Regulatory Purposes*. Benfenati E, ed. Amsterdam, Netherlands:Elsevier, 1–57, https://doi.org/10.1016/B978-044452710-3/50003-3.

Benigni R, Bossa C. 2008a. Predictivity of QSAR. J Chem Inf Model 48(5):971–980, PMID: 18426198, https://doi.org/10.1021/ci8000088.

Benigni R, Bossa C. 2008b. Predictivity and reliability of QSAR models: the case of mutagens and carcinogens. Toxicol Mech Methods 18(2-3):137–147, PMID: 20020910, https://doi.org/10.1080/15376510701857056.

Berhanu WM, Pillai GG, Oliferenko AA, Katritzky AR. 2012. Quantitative structure–activity/property relationships: the ubiquitous links between cause and effect. Chempluschem 77(7):507–517, https://doi.org/10.1002/cplu.201200038.

Boyd DB, March MM. 2006. History of computers in pharmaceutical research and development: a narrative. In: *Computer Applications in Pharmaceutical Research and Development*. Ekins S, ed. New York, NY:Wiley-Interscience, 3–50, https://doi.org/10.1002/0470037237.ch1.

Burello E, Worth AP. 2011. QSAR modeling of nanomaterials. Wiley Interdiscip Rev Nanomed Nanobiotechnol 3(3):298–306, PMID: 21384562, https://doi.org/10.1002/wnan.137.

Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. 2014. QSAR modeling: where have you been? Where are you going to?. J Med Chem 57(12):4977–5010, PMID: 24351051, https://doi.org/10.1021/jm4004285.

Chirico N, Gramatica P. 2011. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model 51(9):2320–2335, PMID: 21800825, https://doi.org/10.1021/ci200211n.

Clark RD, Waldman M. 2012. Lions and tigers and bears, oh my! Three barriers to progress in computer-aided molecular design. J Comput Aided Mol Des 26(1):29–34, PMID: 22160503, https://doi.org/10.1007/s10822-011-9504-3.

Combes RD. 2001. Challenges for computational structure–activity modelling for predicting chemical toxicity: future improvements? Expert Opin Drug Metab Toxicol 7(9):1129–1140, PMID: 21756202, https://doi.org/10.1517/17425255.2011.602066.

Cronin MTD, Jaworska JS, Walker JD, Comber MHI, Watts CD, Worth AP. 2003a. Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environ Health Perspect 111(10):1391–1401, PMID: 12896862, https://doi.org/10.1289/ehp.5760.

Cronin MTD, Walker JD, Jaworska JS, Comber MHI, Watts CD, Worth AP. 2003b. Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. Environ Health Perspect 111(10):1376–1390, PMID: 12896861, https://doi.org/10.1289/ehp.5759.

Cumming JG, Davis AM, Muresan S, Haeberlein M, Chen H. 2013. Chemical predictive modelling to improve compound quality. Nat Rev Drug Discov 12(12):948–962, PMID: 24287782, https://doi.org/10.1038/nrd4128.

Currie LA, Svehla G. 1994. Nomenclature for the presentation of results of chemical analysis (IUPAC Recommendations 1994). Pure Appl Chem 66(3):595–608, https://doi.org/10.1351/pac199466030595.

Das RD, Roy K. 2013. Advances in QSPR/QSTR models of ionic liquids for the design of greener solvents of the futuure. Mol Divers 17(1):151–196, PMID: 23325356, https://doi.org/10.1007/s11030-012-9413-y.

Dearden JC. 2003. In silico prediction of drug toxicity. J Comput Aided Mol Des 17(2-4):119–127, PMID: 13677480, https://doi.org/10.1023/A:1025361621494.

Dearden JC. 2016. The history and development of quantitative structure-activity relationships (QSARs). Int J Quant Struct-Prop Relatsh 1(1):1–44, https://doi.org/10.4018/IJQSPR.2016010101.

Dearden JC. 2017. The history and development of quantitative structure-activity relationships (QSARs): addendum. Int J Quant Struct-Prop Relatsh 2(2):36–46, https://doi.org/10.4018/IJQSPR.2017070104.

Dearden JC, Cronin MTD, Kaiser KLE. 2009. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). SAR QSAR Environ Res 20(3-4):241–266, PMID: 19544191, https://doi.org/10.1080/10629360902949567.

Doweyk AM. 2008. QSAR: dead or alive? J Comput Aided Mol Des 22(2):81–89, PMID: 18189157, https://doi.org/10.1007/s10822-007-9162-7.

Downing J, Murray-Rust P, Tonge AP, Morgan P, Rzepa HS, Cotterill F, et al. 2008. SPECTRa: the deposition and validation of primary chemistry research data in digital repositories. J Chem Inf Model 48(8):1571–1581, PMID: 18661966, https://doi.org/10.1021/ci7004737.

Editorial. 2006. QSAR/QSPR and proprietary data. J Med Chem 49:3431–3431, https://doi.org/10.1021/jm068019l.

EPCEU (European Parliament and the Council of the European Union). 2006. Regulation (EC) No. 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No. 793/93 and Commission Regulation (EC) No. 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. Official Journal of the European Union 396. http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02006R1907-20161011&from=EN [accessed 20 August 2017].

EPCEU. 2009. Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on Cosmetic Products. http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32009R1223 [accessed 2 June 2018].

Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P. 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 111(10):1361–1375, PMID: 12896860, https://doi.org/10.1289/ehp.5758.

Fourches D, Muratov E, Tropsha A. 2010. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 50(7):1189–1204, PMID: 20572635, https://doi.org/10.1021/ci100176x.

Fourches D, Muratov E, Tropsha A. 2016. Trust, but verify II: a practical guide to chemogenomics data curation. J Chem Inf Model 56(7):1243–1252, PMID: 27280890, https://doi.org/10.1021/acs.jcim.6b00129.

Fujita T, Winkler DA. 2016. Understanding the roles of the "two QSARs." J Chem Inf Model 56(2):269–274, PMID: 26754147, https://doi.org/10.1021/acs.jcim.5b00229.

Gallegos Salinger A, Tsakovska I, Pavan M, Patlewicz G, Worth AP. 2007. Evaluation of SARs for the prediction of skin irritation/corrosion potential: structural inclusion rules in the BfR decision support system. SAR QSAR Environ Res 18(3-4):331–342, PMID: 17514574, https://doi.org/10.1080/10629360701304014.

Gedeck P, Kramer C, Ertl P. 2010. Computational analysis of structure–activity relationships. Prog Med Chem 49:113–160, PMID: 20855040, https://doi.org/10.1016/S0079-6468(10)49004-9.

Golbraikh A, Tropsha A. 2002. Beware of q2! J Mol Graph Model 20(4):269–276, PMID: 11858635, https://doi.org/10.1016/S1093-3263(01)00123-1.

Gramatica P. 2007. Principles of QSAR models validation: internal and external. QSAR Comb Sci 26(5):694–701, https://doi.org/10.1002/qsar.200610151.

Hansch C, Fujita T. 1964. p-σ-π analysis. A method for the correlation of biological activity and chemical structure. J Am Chem Soc 86(8):1616–1626, https://doi.org/10.1021/ja01062a035.

Hansch C, Hoekman D, Gao H. 1996. Comparative QSAR: toward a deeper understanding of chemicobiological interactions. Chem Rev 96(3):1045–1076, PMID: 11848780, https://doi.org/10.1021/cr9400976.

Hansch C, Leo A, Taft RW. 1991. A survey of Hammett substituent constants and resonance and field parameters. Chem Rev 91(2):165–195, https://doi.org/10.1021/cr00002a004.

Hansch C, Hoekman D, Leo A, Weininger D, Selassie CD. 2002. Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology. Chem Rev 102(3):83–812, PMID: 11890757, https://doi.org/10.1021/cr0102009.

Hartung T, Bremer S, Casati S, Coecke S, Corvi R, Fortaner S, et al. 2004. A modular approach to the ECVAM principles on test validity. Altern Lab Anim 32(5):467–472, PMID: 15656771.

Huang J, Fan X. 2011. Why QSAR fails: an empirical evaluation using conventional computational approach. Mol Pharm 8(2):600–608, PMID: 21370915, https://doi.org/10.1021/mp100423u.

Johnson SR. 2007. The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). J Chem Inf Model 48(1):25–26, PMID: 18161959, https://doi.org/10.1021/ci700332k.

Jorgensen WL. 2006. QSAR/QSPR and proprietary data. J Chem Inf Model 46(3):937, https://doi.org/10.1021/ci0680079.

JRC QSAR Model Database. User Manual, Version 2, European Commission, DG Joint Research Centre, Institute for Health and Consumer Protection, Systems Toxicology Unit. https://qsardb.jrc.ec.europa.eu/qmrf [accessed 30 June 2017].

Käärik M, Maran U, Arulepp M, Perkson A, Leis J. 2018. Quantitative nanostructure–property relationships for the nanoporous carbon: predicting the performance of energy storage materials. ACS Appl Energy Mater 1(8):4016–4024, https://doi.org/10.1021/acsaem.8b00708.

Katritzky AR, Kuanar M, Slavov S, Hall CD, Karelson M, Kahn I, et al. 2010. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. Chem Rev 110(10):5714–5789, https://doi.org/10.1021/cr900238d.

Katritzky AR, Maran U, Lobanov VS, Karelson M. 2000. Structurally diverse quantitative structure–property relationship correlations of technologically relevant physical properties. J Chem Inf Comput Sci 40(1):1–18, PMID: 10661545, https://doi.org/10.1021/ci9903206.

Kruhlak NL, Contrera JF, Benz RD, Matthews EJ. 2007. Progress in QSAR toxicity screening of pharmaceutical impurities and other FDA regulated products. Adv Drug Deliv Rev 59(1):43–55, PMID: 17229485, https://doi.org/10.1016/j.addr.2006.10.008.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res 19(9):1639–1645, PMID: 19541911, https://doi.org/10.1101/gr.092759.109.

Le T, Epa VC, Burden FR, Winkler DA. 2012. Quantitative structure–property relationship modeling of diverse materials properties. Chem Rev 112(5):2889–2919, PMID: 22251444, https://doi.org/10.1021/cr200066h.

Maran U, Sild S, Mazzatorta P, Casalegno M, Benfenati E, Romberg M. 2007. Grid computing for the estimation of toxicity: acute toxicity on fathead minnow (*Pimephales promelas*). In: *Distributed, High-Performance and Grid Computing in Computational Biology*, Vol. 4360. Dubitzky W, Schuster A, Sloot PMA, Schhroeder M, Romberg M, eds. Berlin Heidelberg:Springer-Verlag, 60–74, https://doi.org/10.1007/978-3-540-69968-2_6.

Martin TM, Harten P, Young DM, Muratov EN, Golbraikh A, Zhu H, et al. 2012. Does rational selection of training and test sets improve the outcome of QSAR modeling? J Chem Inf Model 52(10):2570–2578, PMID: 23030316, https://doi.org/10.1021/ci300338w.

Muratov EN, Varlamova EV, Artemenko AG, Polishchuk PG, Kuz'min VE. 2012. Existing and developing approaches for QSAR analysis of mixtures. Mol Inform 31(3-4):202–221, PMID: 27477092, https://doi.org/10.1002/minf.201100129.

Nendza M, Russom CL. 1991. QSAR modelling of the ERL-D fathead minnow acute toxicity database. Xenobiotica 21(2):147–170, PMID: 2058173, https://doi.org/10.3109/00498259109039458.

Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, et al. 2005. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. Altern Lab Anim 33(2):155–173, PMID: 16180989.

Nicolotti O, Benfenati E, Carotti A, Gadaleta D, Gissi A, Mangiatordi GF, et al. 2014. REACH and in silico methods: an attractive opportunity for medicinal chemists. Drug Discov Today 19(11):1757–1768, PMID: 24998783, https://doi.org/10.1016/j.drudis.2014.06.027.

OECD (Organization for Economic Cooperation and Development). 2007. "Environment Directorate: Guidance Document on the Validation of (Q)SAR Models." OECD Environment Health and Safety Publications Series on Testing and Assessment, No. 69, ENV/JM/MONO(2007)2. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2 [accessed 2 June 2018].

Patlewicz G, Fitzpatrick JM. 2016. Current and future perspectives on the development, evaluation, and application of in silico approaches for predicting toxicity. Chem Res Toxicol 29(4):438–451, PMID: 26686752, https://doi.org/10.1021/acs.chemrestox.5b00388.

Piir G, Sild S, Maran U. 2014. Classifying bio-concentration factor with random forest algorithm, influence of the bio-accumulative vs. non-bio-accumulative compound ratio to modelling result, and applicability domain for random forest model. SAR QSAR Environ Res 25(12):967–981, PMID: 25482723, https://doi.org/10.1080/1062936X.2014.969310.

Price NR, Watkins RW. 2003. Quantitative structure-activity relationships (QSAR) in predicting the environmental safety of pesticides. Pest Outlook 14(3):127–129, https://doi.org/10.1039/b305506j.

QSARDB Repository. http://qsardb.org/ [accessed 2 June 2018].

Roy PP, Kovarich S, Gramatica P. 2011. QSAR model reproducibility and applicability: a case study of rate constants of hydroxyl radical reaction models applied to polybrominated diphenyl ethers and (benzo-)triazoles. J Comput Chem 32(11):2386–2396, PMID: 21541967, https://doi.org/10.1002/jcc.21820.

Ruusmann V, Maran U. 2013. From data point timelines to a well curated data set, data mining of experimental data and chemical structure data from scientific articles, problems and possible solutions. J Comput Aided Mol Des 27(7):583–603, PMID: 23884706, https://doi.org/10.1007/s10822-013-9664-4.

Ruusmann V, Sild S, Maran U. 2014. QSAR DataBank - an approach for the digital organization and archiving of QSAR model information. J Cheminform 6:25, PMID: 24910716, https://doi.org/10.1186/1758-2946-6-25.

Ruusmann V, Sild S, Maran U. 2015. QSAR DataBank repository: open and linked qualitative and quantitative structure–activity relationship models. J Cheminform 7:32, PMID: 26110025, https://doi.org/10.1186/s13321-015-0082-6.

Scior T, Medina-Franco JL, Do QT, Martinez-Mayorga K, Yunes Rojas JA, Bernard P. 2009. How to recognize and workaround pitfalls in QSAR studies: a critical review. Curr Med Chem 16(32):4297–4313, PMID: 19754417, https://doi.org/10.2174/092986709789578213.

SCOPUS. Scopus literature database. https://www.scopus.com/ [accessed 30 June 2017].

Seddon G, Lounnas V, McGuire R, van den Bergh T, Bywater RP, Oliveira L, et al. 2012. Drug design for ever, from hype to hope. J Comput Aided Mol Des 26(1):137–150, PMID: 22252446, https://doi.org/10.1007/s10822-011-9519-9.

Sild S, Maran U, Lomaka A, Karelson K. 2006. Open computing grid for molecular science and engineering. J Chem Inf Model 46(3):953–959, PMID: 16711713, https://doi.org/10.1021/ci050354f.

Sild S, Piir G, Neagu D, Maran U. In press. Storing and using quantitative and qualitative structure–activity relationships in the era of toxicological and chemical data expansion. In: *Big Data in Predictive Toxicology*. Neagu D, Richarz A, eds. Cambridge:Royal Society of Chemistry.

Stouch TR, Kenyon JR, Johnson SR, Chen XQ, Doweyko A, Li Y. 2003. In silico ADME/Tox: why models fail. J Comput Aided Mol Des 17(2-4):83–92, PMID: 13677477, https://doi.org/10.1023/A:1025358319677.

Tantra R, Oksel C, Puzyn T, Wang J, Robinson KN, Wang XZ, et al. 2015. Nano(Q)SAR: challenges, pitfalls and perspectives. Nanotoxicology 9(5):636–642, PMID: 25211549, https://doi.org/10.3109/17435390.2014.952698.

Tetko IV, Bruneau P, Mewes HW, Rohrer DC, Poda GI. 2006. Can we estimate the accuracy of ADME–Tox predictions? Drug Discov Today 11(15-16):700–707, PMID: 16846797, https://doi.org/10.1016/j.drudis.2006.06.013.

Tetko IV, Engkvist O, Koch U, Reymond JL, Chen H. 2016. BIGCHEM: challenges and opportunities for big data analysis in chemistry. Mol Inform 35(11-12):615–621, PMID: 27464907, https://doi.org/10.1002/minf.201600073.

Tetko IV, Maran U, Tropsha A. 2017. Public (Q)SAR services, integrated modeling environments, and molecular repositories on the web: state of the art and perspectives for future development. Mol Inf 36(3):1600082, PMID: 27778468, https://doi.org/10.1002/minf.201600082.

Tropsha A. 2010. Best practices for QSAR model development, validation, and exploitation. Mol Inform 29(6-7):476–488, PMID: 27463326, https://doi.org/10.1002/minf.201000061.

Tropsha A, Gramatica P, Gombar VK. 2003. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 22(1):69–77, https://doi.org/10.1002/qsar.200390007.

Tsakovska I, Gallegos Salinger A, Netzeva T, Pavan M, Worth AP. 2007. Evaluation of SARs for the prediction of eye irritation/corrosion potential: structural inclusion rules in the BfR decision support system. SAR QSAR Environ Res 18(3-4):221–235, PMID: 17514567, https://doi.org/10.1080/10629360701304063.

Winkler DA, Burden FR, Yan B, Weissleder R, Tassa C, Shaw S, et al. 2014. Modelling and predicting the biological effects of nanomaterials. SAR QSAR Environ Res 25(2):161–172, PMID: 24625316, https://doi.org/10.1080/1062936X.2013.874367.

WOS. Web of Science. https://webofknowledge.com/ [accessed 30 June 2017].